An empirical likelihood approach under cluster sampling with missing observations

Yves G. Berger

Received: date / Revised: date

Abstract The parameter of interest considered is the unique solution to a set of estimating equations, such as regression parameters of generalised linear models. We consider a design-based approach; that is, the sampling distribution is specified by stratification, cluster (multi-stage) sampling, unequal selection probabilities, side information and a response mechanism. The proposed empirical likelihood approach takes into account of these features. Empirical likelihood has been mostly developed under more restrictive settings, such as independent and identically distributed assumption, which is violated under a design-based framework. A proper empirical likelihood approach which deals with cluster sampling, missing data and multidimensional parameters is absent in the literature. This paper shows that a cluster-level empirical log-likelihood ratio statistic is pivotal. The main contribution of the paper is to provide the rigorous asymptotic theory and underlining regularity conditions which imply $\sqrt{n}$-consistency and the Wilks's theorem or self-normalisation property. Negligible and large sampling fractions are considered.

Keywords design-based approach · estimating equations · response mechanism · response propensities · stratification · side information · unequal probabilities

1 Introduction

We consider that the sample data are selected with a stratified cluster (multistage) sampling design with unequal probabilities. Side information is also

Yves G. Berger
University of Southampton,
Southampton Statistical Sciences Research Institute,
Southampton, SO17 1BJ. United Kingdom
E-mail: Y.G.Berger@soton.ac.uk
URL: http://www.yvesberger.co.uk

included. Empirical likelihood approaches for missing data have been developed under more restrictive settings, which do not take into account of the complexity of the design and parameters. For example, Qin et al's (2009) empirical likelihood approach for missing data, does not includes cluster sampling, unequal probabilities and population level information. The complexity of the sampling design is the primarily focus of this paper. Standard empirical likelihood approach based on the complete case cannot be straightforwardly implemented, because it would not take into account of the sampling design and response mechanism. We proposed a general empirical likelihood approach which accommodates these complex features. It naturally includes adjustments for missing data. In §6, we give regularity conditions which imply the Wilks's theorem and $\sqrt{n}$-consistency.

Parameters are often multidimensional, such as parameters of generalised linear models. However, most of the design-based literature on missing data deals with unidimensional parameters, such as totals, means, ratios and quantiles (e.g. Haziza and Lesage 2016). However, in the presence of non-response, the parameter is multidimensional, because the parameter of interest depends on non-response parameters. For example, if we wish to estimate a mean and we have c re-weighting classes. We have a multidimensional parameter of size $c + 1$ containing the mean and c response probabilities, one for each class. It is common practice to treat the estimated response probabilities as deterministic, which may affect confidence intervals (Valliant 2004). In their simplest forms, these estimated response probabilities reduces to response rates. For example, when all the units have the same response probability, we have one re-weighting class and an estimator of this probability is the overall response rate. It is common practice to ignore the estimation of these probabilities and to treat them as if they were deterministic. The randomness of these probabilities is taken into account within the empirical likelihood confidence intervals proposed.

Pseudo-likelihood (Binder 1983) can be used for regression parameters. However, pseudo-likelihood confidence intervals are based on Wald's statistics, involving linearisation. There is no Wilks's type theorem for pseudo-likelihood. Empirical likelihood has the advantage of having data driven and range preserving confidence intervals, based on a self-normalising empirical log-likelihood ratio statistic. The empirical likelihood approach proposed may provide better confidence intervals than those based on Wald's type statistics.

The mainstream empirical likelihood theory under independent and identically distributed (i.i.d.) observations, was developed by Owen (1988) and Qin and Lawless (1994). Wang and Rao (2002a) proposed several empirical likelihood approaches for imputed estimators of means, under a i.i.d. setting. It has been extended for linear models and estimating equations by Wang and Rao (2002b); Wang and Chen (2009); Qin et al (2009). We consider a different situation when we have a stratified cluster sampling design with unequal probabilities and side information.

Survey data are often clustered; that is, the population frame is split into small groups of units, called clusters. A specified number of clusters are sampled. Units are selected within each cluster sampled. This is a widely used technique for social surveys. We have a single-stage design when all the units are selected within each cluster sampled. In both cases, the observations are not i.i.d., and the customary empirical likelihood approach (Owen 2001) cannot be straightforwardly extended. We shall see that this customary empirical likelihood approach based on the completed cases produces confidence intervals with coverages significantly different from the nominal value.

Chen and Sitter's (1999) pseudo-empirical likelihood is based on a weighted empirical likelihood function. It was extended for stratified simple sample with missing data by Fang et al (2009, 2010). The pseudo-empirical log-likelihood ratio statistic is not pivotal (Wu and Rao 2006). Hence, the self-normalisation property does not hold. Confidence intervals can still be obtained using linearisation or by adjusting the pseudoempirical log-likelihood ratio statistic by a ratio of variance estimates, as in Wu and Rao (2006) and Wang and Rao (2002b). This adjustment is limited to unidimensional parameters, and cannot be used with multidimensional parameters. Chen and Kim's (2014) population empirical likelihood approach is based on single-stage Poisson sampling with random sample size, which is not considered in this paper, because we consider that the number of clusters sampled is deterministic.

Berger and Torres (2016) extended Owen's (1988) approach for single-stage unequal probabilities sampling and full response, when we have a single estimating equation and fixed sample size. Oğuz-Alper and Berger (2016) generalised this approach for multidimensional parameters under full response. The fact that the self-normalisation property holds is major advantage over the pseudo-empirical likelihood approach. A comparison between Oğuz-Alper and Berger's (2016) approach and pseudo-empirical likelihood can be found in Berger (2018). Berger and Torres (2016) proposed an extension for multi-stage design and a single estimating equation. They conjectured that the empirical log-likelihood ratio statistic is pivotal, under full response. In this paper, we proof this conjecture, under a more general setting involving missing data and multidimensional parameters. The contribution of this paper is to provide regularity conditions on the design which ensures that the profile cluster-level empirical log-likelihood ratio statistic is pivotal. The primarily focus of this paper is the complexity of the sampling design, rather than missingness. The former is add-on feature which cannot be avoided with clustered samples. Imputation is not covered and is beyond the scope of this paper.

The aim of this paper is to develop a rigorous asymptotic theory of empirical likelihood under multistage designs and nonresponse. We shall use response propensities to adjust for missing data, as in Qin et al (2009), but we consider a different setting when we have a stratified cluster sampling design with unequal probabilities and side information. We show that the independence between the response mechanism and the sampling design implies that the empirical log-likelihood ratio statistic is pivotal and does not need to be adjusted for missing data. The empirical log-likelihood ratio statistic takes the response mechanism into account. First, we show that this result holds for negligible sampling fractions. Then, we show how the empirical log-likelihood ratio statistic can be adjusted for large sampling fractions. We provide the regularity conditions on the multistage sampling design and response mechanism which ensure that the empirical likelihood estimator is consistent and that the empirical log-likelihood ratio statistic is pivotal.

It is common practice to treat the estimated response propensities as deterministic within variance estimators. This may shortened the confidence intervals (Valliant 2004). We show that the empirical log-likelihood ratio statistic possesses the self-normalising property, while taking into account of the estimation of these propensities. This allows confidence intervals which reflect the estimation of these propensities.

Inverse probability weighting approaches for handling missing data is well developed in the survey sampling literature (e.g. Brick and Kalton 1996; Brick and Montaquila 2009). Most of them are based on a propensity model, with stochastic response as a second phase (Särndal and Swensson 1987), with unknown response propensities. Non-response bias reduction can be achieved under accurate estimation of the response propensities. The propensity model used is often a logistic model containing categories or classes, as in Little (1986). We shall use a similar approach. Another approach is a non-response weighting adjustment based on calibration (Särndal and Lundström 2005), based on auxiliary information at sample or population level (Brick and Kalton 1996; Lundström and Särndal 1999). Other technique involves calculating an upper bound for the non-response bias (Montaquila et al 2008).

There are three main approaches for variance estimation: Jackknife (Rao and Shao 1992; Berger and Rao 2006), bootstrap (Kovar et al 1988; Rust and Rao 1996) and linearisation (Wolter 2007; Binder 1983; Deville 1999). Asymptotic theory of bootstrap is restricted to simple settings. Its properties is often limited to means, and solely based on simulations. Valliant (2004) compared these approaches via simulation. Brick and Montaquila (2009) pointed out that more research is needed on the effect of non-response weighting on confidence intervals. We proposed to fill this gap, by showing how this effect can be taken into account, by using a profile empirical log-likelihood ratio statistic. Its implementation does not involves variance estimation, re-sampling or linearisation.

In §2, we define the response mechanism and the sampling design. The class of multi-stage designs is defined in §3. The parameters of interest and side information are defined in §4. In §5, we describe the empirical likelihood approach proposed. The asymptotic results can be found in §6. The key results is Theorem 1, which shows that the empirical log-likelihood ratio statistic is pivotal. The approach proposed is extended for two-stage designs with large sampling fraction in §7. The proofs are given in in the online supplement. Simulation results, found in §8, show that the approach proposed is robust against skewed data, extreme values, and large sampling fraction. An example of application to real survey data can be found in §9.

2 Response mechanism, sampling design and sample data

Consider a population $\mathcal{U} = \{1, \ldots, \mathcal{N}\}$ containing $\mathcal{N}$ units. Let

$$\begin{aligned}\boldsymbol{\xi} &:= (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i, \ldots, \boldsymbol{\xi}_{\mathcal{N}})^\top,\\ \boldsymbol{\zeta} &:= (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_i, \ldots, \boldsymbol{\zeta}_{\mathcal{N}})^\top,\\ \boldsymbol{y} &:= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_{\mathcal{N}})^\top,\end{aligned}$$

where $\boldsymbol{\xi}_i \in \mathbb{R}^{d_\xi}$, $\boldsymbol{\zeta}_i \in \mathbb{R}^{d_\zeta}$ and $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$ denote vectors of constant values attached to unit $i \in \mathcal{U}$. The vectors $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$ and $\boldsymbol{y}$ are called respectively the '*non-response variables*', the '*design variables*' and the '*variables of interest*'. We consider that some components of $\boldsymbol{y}$ are subject to missingness. The variables $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are not subject to missingness.

We consider a '*design-based approach*' (Neyman 1938); that is, $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$ and $\boldsymbol{y}$ are treated as constants. This is a non-parametric approach, because we do no assume any distribution for $\boldsymbol{\xi}_i$,$\boldsymbol{\zeta}_i$ and $\boldsymbol{y}_i$. The design-based approach is different from mainstream statistics and is the core of survey data estimation. The response mechanism and the sampling design are respectively defined in §2.1 and 2.2. The response mechanism specifies the random process which determines which unit i is missing (Rubin 1976). The sampling design characterises the random selection of a sample within $\mathcal{U}$. We assume that the response mechanism and the sampling design are independent random processes.

2.1 Response mechanism

Let r_i denotes the response indicator of $i \in \mathcal{U}$, where $r_i = 1$ if i is not missing and $r_i = 0$ if i is missing. The response mechanism is characterised by the probability space $\{\Omega_r, \sigma(\Omega_r), \mathbb{P}_r\}$, where

$$\Omega_r = \{\boldsymbol{r} := (r_1, \ldots, r_i, \ldots, r_{\mathcal{N}})^\top : r_i = 0 \text{ or } 1\}, \tag{1}$$

generates the σ-algebra $\sigma(\Omega_r)$. The probability $\mathbb{P}_r : \sigma(\Omega_r) \to [0, 1]$ is given by

$$\mathbb{P}_r(\boldsymbol{r}, \boldsymbol{\xi}, \boldsymbol{\lambda}_0) := \prod_{i \in U} P_i(\boldsymbol{\lambda}_0)^{r_i}\{1 - P_i(\boldsymbol{\lambda}_0)\}^{1-r_i}, \tag{2}$$

where

$$P_i(\boldsymbol{\lambda}) := \mathcal{F}^{-1}(\boldsymbol{\xi}_i^\top \boldsymbol{\lambda}), \tag{3}$$

where $\mathcal{F}^{-1} : \mathbb{R} \to (0, 1]$ is the inverse of a link function $\mathcal{F}$ (e.g. logit, probit, complementary log-log). Definition (2) means that $r_i \sim \text{Bernoulli}(\rho_i)$, with $r_i \perp\!\!\!\perp r_j$ for $i \neq j$ and

$$\rho_i := P_i(\boldsymbol{\lambda}_0)\cdot \tag{4}$$

The quantity ρ_i is called the '*response propensity*' of unit $i \in \mathcal{U}$ and $\boldsymbol{\lambda}_0 \in \mathbb{R}^{d_\xi}$ is called the '*response parameter*'. The $\boldsymbol{\xi}_i$ denotes variables that explains the missingness. For example, the $\boldsymbol{\xi}_i$ may contain some geographical variables or variables available in a population register or census. We shall assume that the ρ_i are unknown and correctly specified by (3).

The choice of $\boldsymbol{\xi}_i$ is discussed in Kalton (1983), Särndal and Lundström (2005), Little and Vartivarian (2005). In practice, it is preferable to have $\boldsymbol{\xi}_i$ being a set of dichotomous variables representing re-weighting classes, with uniform response propensities within classes (Little 1986; Haziza and Beaumont 2007; Brick and Montaquila 2009), since logistic model with continuous variable may give unstable estimates (Little 1986). Little (1986) proposed to create quantile classes from fitted response probabilities, and use them in a logistic model to obtain more stable propensities. In this case, $\boldsymbol{\xi}_i$ is a matrix of dummy variables specifying the classes, and $\boldsymbol{\lambda}_0$ is a vector containing the response rates for each classes.

2.2 Sampling design

A sample is a collection of units from $\mathcal{U}$. A sample is not necessarily a subset of $\mathcal{U}$, because the same unit can be sampled several times, under with replacement sampling. Let d_i denote the number of time a unit $i \in \mathcal{U}$ is selected, with $d_i = 0$ when the unit i is not selected. The sample size is $\nu := \sum_{i\in\mathcal{U}} d_i$. Consider the set Ω_d of all possible samples:

$$\Omega_d = \{\boldsymbol{d} := (d_1, \ldots, d_i, \ldots, d_{\mathcal{N}})^\top : d_i \in \mathbb{N}\}\cdot$$

We consider the probability space $\{\Omega_d, \sigma(\Omega_d), \mathbb{P}_d\}$, where $\mathbb{P}_d : \sigma(\Omega_d) \to [0, 1]$ is a probability measure called '*sampling design*'. This probability is denoted by $\mathbb{P}_d(\boldsymbol{d}, \boldsymbol{\zeta})$ and is a function of design variables $\boldsymbol{\zeta}$, which includes information about strata, clusters or selection probabilities. The class of sampling designs considered is defined in §3.

2.3 Product space and sample data

The key assumption is the independence between the response mechanism and the sampling design. This is a weak assumption often met in practice. It means that ρ_i does not depend on the sample selected. Thus, we have the product probability space $\{\Omega_r \times \Omega_d, \sigma(\Omega_r) \otimes \sigma(\Omega_d), \mathbb{P}_{r,d}\}$, where $\mathbb{P}_{r,d}(\boldsymbol{r}, \boldsymbol{d}, \boldsymbol{\xi}, \boldsymbol{\lambda}_0, \boldsymbol{\zeta}) = \mathbb{P}_r(\boldsymbol{r}, \boldsymbol{\xi}, \boldsymbol{\lambda}_0) \times \mathbb{P}_d(\boldsymbol{d}, \boldsymbol{\zeta})$. A random variable is a real measurable function on that product probability space.

An outcome of $\Omega_r \times \Omega_d$ is $\omega_{r,d} := \{(r_i, d_i)^\top : i \in U\}$, with $\boldsymbol{\xi}_i$, $\boldsymbol{\zeta}_i$ and $\boldsymbol{y}_i$ being associated to $i \in U$. We adopt the convention that $\boldsymbol{\xi}_i$, $\boldsymbol{\zeta}_i$ and $\boldsymbol{y}_i$ are only known for i such that $d_i \neq 0$, with some components of $\boldsymbol{y}_i$ missing when $r_i = 0$. Instead of the outcome $\omega_{r,d}$, we prefer to use the equivalent concept of '*sample data*' given by

$$\mathscr{D}_{r,d} := \left\{(r_i, d_i, \boldsymbol{\xi}_i^\top, \boldsymbol{\zeta}_i^\top, \boldsymbol{y}_i^\top)^\top : i \in \mathcal{U}, d_i \neq 0\right\},$$

because the variables $\boldsymbol{\xi}_i$, $\boldsymbol{\zeta}_i$ and $\boldsymbol{y}_i$ are vectors of constants. A real Borel-measurable function of $\mathscr{D}_{r,d}$ is a random variable. For example, the sample mean of the non-missing values, $(\sum_{i \in \mathcal{U}} d_i r_i)^{-1} \sum_{i \in \mathcal{U}} d_i r_i \boldsymbol{y}_i$, is a random variable with a sampling distribution specified by $\mathbb{P}_{r,d}$. Note that this random variable is usually a biased estimator of the population mean of $\boldsymbol{y}_i$, because it does not contains non-response adjustments.

The design-based framework described in this § is different from mainstream statistics, because the $\boldsymbol{\xi}_i$, $\boldsymbol{\zeta}_i$ and $\boldsymbol{y}_i$ are constants. The sampling distribution is specified by probability space $\{\Omega_r \times \Omega_d, \sigma(\Omega_r) \otimes \sigma(\Omega_d), \mathbb{P}_{r,d}\}$ or equivalently, by the response mechanism and the sampling design. The advantage is that it is not necessary to specify a distribution for the variable of interest, and it provides robust non-parametric approach for estimation.

2.4 Some remarks

Response mechanisms are often classified as '*missing completely at random*' (MCAR), '*missing at random*' (MAR) and '*not missing at random*' (NMAR) (Rubin 1976; Little and Rubin 2002). We have a MCAR mechanisms, when there is no correlation between $\boldsymbol{\xi}$ and $\boldsymbol{y}$. We should not view this correlation from a probabilist point of view, but simply as the descriptive correlation measured between the $\mathcal{N}$-vectors of constants within $\boldsymbol{\xi}$ and $\boldsymbol{y}$. We have a NMAR mechanism, when $\boldsymbol{\xi}$ and $\boldsymbol{y}$ have common variables or are correlated. We do not consider NMAR mechanisms. We assume that the response mechanism is MAR. In §4, we will see that the MAR assumption is linked with the estimating function. In §8.2, a simulation study evaluates the approach proposed under a NMAR mechanism.

Even under MAR, there is still a non-response component within the variance. Response propensities need to be taken into account within the weights to reflect the fact that the complete case sample is smaller that the sample selected. Ignoring the response mechanism under-estimates the variance. Our profile empirical log-likelihood ratio statistic takes the response mechanism into account within confidence intervals and p-values. Since we assume that the response mechanism is independent of the sampling design, we can consider that non-response occurs before sampling as in Fay (1991) and Shao and Steel (1999). This allows to have the effect of the design and non-response included within a single term, which is captured by the empirical log-likelihood ratio statistic. Since, non-response is stochastic, the fact that it occurs before or after sampling, has no implication for the expectation and variance of point estimates (Fay 1991).

Equation (2) implicitly assumes that $\boldsymbol{\xi}$ explains missingness. In §5, response propensities will be used to adjust for missing data. Even under MAR, the response mechanism needs to be taken into account for propensity weighting. Within the variance, there is also a component due to missingness, to reflect the loss of efficiency that occurs because of the non-response mechanism.

The sampling design is informative when some variables of $\boldsymbol{\xi}$ are correlated with $\boldsymbol{\zeta}$ (Pfeffermann et al 1998). Ignoring informativeness may result in invalid inference. Informativeness will be taken into account by incorporating $\boldsymbol{\zeta}$ within the estimating equations (see (7)). We assume that the design variables $\boldsymbol{\zeta}$ are known for all the sampled units.

3 Cluster sampling design

In practice, sampling designs commonly used, involves stratification, clusters and unequal selection probabilities (e.g. Brewer and Gregoire 2009). In this §, we define the class of sampling designs considered.

Suppose that the population $\mathcal{U}$ is split into N non-overlapping subsets $\widetilde{U}_k$ called clusters, where $k \in \widetilde{U} = \{1, \ldots, N\}$, $\widetilde{U}$ denotes the population of N clusters and $\cup_{k\in\widetilde{U}}\widetilde{U}_k = \mathcal{U}$. Note that N is different from the population size $\mathcal{N}$. We assume that $\widetilde{U}$ is split into H non-overlapping strata, $\widetilde{U}_1, \ldots, \widetilde{U}_H$, such that $\cup_{h=1}^{H}\widetilde{U}_h = \widetilde{U}$. We assume that a sample $\widetilde{\boldsymbol{S}}_h$ of n_h clusters is selected independently with-replacement within $\widetilde{U}_h$, with probabilities π_k/n_h, where $\sum_{k\in\widetilde{U}_h}\pi_k = n_h$. The sample $\widetilde{\boldsymbol{S}}_h$ contains n_h clusters' labels selected after n_h successive draws. The overall sample of PSU's is denoted $\widetilde{\boldsymbol{S}} = \cup_{h=1}^{H}\widetilde{\boldsymbol{S}}_h$ and contains $n = \sum_{h=1}^{H} n_h$ cluster labels. An important feature of this design is that the clusters are selected with unequal probabilities.

Within each cluster $\widetilde{U}_k$ sampled, we select a without replacement sample $\boldsymbol{S}_k$ of ν_k units. Let $\pi_{i|k}$ denote the conditional inclusion probability of a unit i in $\widetilde{U}_k$. Any sampling designs can be used to select $\boldsymbol{S}_k$. The final sample $\boldsymbol{S} = \cup_{k\in\widetilde{\boldsymbol{S}}}\boldsymbol{S}_k$ contains $\nu = \sum_{k\in\widetilde{\boldsymbol{S}}}\nu_k$ units. We have a single-stage sampling design when we select all the units of each cluster sampled; that is, $\boldsymbol{S}_k = \widetilde{U}_k$.

The sample $\boldsymbol{S}$ contains the labels of units selected, some of them can appear several times within $\boldsymbol{S}$. There is a bijection between all possible samples $\boldsymbol{S}$ and Ω_d, because each $\boldsymbol{S}$ can be paired with a single $\boldsymbol{d} \subset \Omega_d$. Thus, the probability space $\{\Omega_d, \sigma(\Omega_d), \mathbb{P}_d\}$ also describe the random, selection of the sample $\boldsymbol{S}$.

The design variables $\boldsymbol{\zeta}$ specify the clusters, the stratification, the π_k and the $\pi_{i|k}$. We assume that the design variables $\boldsymbol{\zeta}$ are known for the sampled units. However, these variables may be not available to survey data users. Most of the standard survey sampling literatures (e.g. Särndal et al 1992; Wolter 2007) rely on this assumption. Exact analytic approaches for variance estimation are not possible without this information. Proxies need to be used when some of these variables are not available. For example, with the "*European Union Statistics on Income and Living Conditions*" (EU-SILC) survey (Eurostat 2012), the π_k are available, geographical variables can be used as proxies for stratification and survey weights can used within the sums where $\pi_{i|k}$ is needed (see (10) and (11)). Details on how to create these proxies can be found in Osier et al (2013). We also have an example in §9.

4 Parameters of interest and side information

The parameter of interest $\boldsymbol{\tau}_0$ is a function of $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$ and $\boldsymbol{y}$, which is the solution to p estimating equations,

$$\boldsymbol{G}(\boldsymbol{\tau}) := \sum_{i \in \mathcal{U}} \boldsymbol{g}_i(\boldsymbol{\tau}) = \mathbf{0}_p, \tag{5}$$

where $\boldsymbol{g}_i(\boldsymbol{\tau}) = \boldsymbol{g}(\boldsymbol{\tau}, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i, \boldsymbol{y}_i) \in \mathbb{R}^p$ is an estimating function of $\boldsymbol{\tau}$, $\boldsymbol{\xi}_i$, $\boldsymbol{\zeta}_i$ and $\boldsymbol{y}_i$. Here, $\boldsymbol{\tau}_0 \in \boldsymbol{\mathcal{T}} \subset \mathbb{R}^{p'}$ and $\boldsymbol{\tau} \in \boldsymbol{\mathcal{T}}$, where $\boldsymbol{\mathcal{T}}$ is compact and $p' \leqslant p$. The vector $\mathbf{0}_p$ is the p-vector of zeros. We assume that the solution to (5) is unique. For example, $\boldsymbol{\tau}_0$ can be a vector of population regression coefficient of a generalised linear regression model (e.g. Binder and Patak 1994; Chen and Van Keilegom 2009). Examples of logistic and poisson regression parameters can be found in §8.

Asymptotically unbiased estimation of $\boldsymbol{G}(\boldsymbol{\tau})$ is the key aspect of the theory of estimating equation in survey sampling (e.g. Godambe and Thompson 2009). In order for weighted estimator of $\boldsymbol{G}(\boldsymbol{\tau})$ to be unbiased, the response mechanism must be such that there is no correlation between $\boldsymbol{g}_i(\boldsymbol{\tau})$ and ρ_i given by (4). This derived from the standard theory of weighted estimator of totals (e.g. Haziza and Beaumont 2007; Haziza 2009). When $\boldsymbol{g}_i(\boldsymbol{\tau})$ is a non-linear function of $\boldsymbol{y}$, the covariance between $\boldsymbol{g}_i(\boldsymbol{\tau})$ and ρ_i could be negligible, even if $\boldsymbol{y}$ and ρ_i are dependent, under a non-ignorable (NMAR) response mechanism. Hence, ignorability of the response mechanism depends on the estimating function $\boldsymbol{g}_i(\boldsymbol{\tau})$ or the parameter to estimate. An example can be found in §8.2.

We may have some '*side information*' in the form of population-level means, counts or proportions from large external censuses or surveys, known without sampling errors as in (Owen 2001 §3.10). In the econometric literature, this is known as '*deterministic macro-level information*' or '*exact knowledge*' (Imbens and Lancaster 1994). In other words, we assume that we know a vector $\boldsymbol{\varphi}_0 \in \boldsymbol{\Phi} \subset \mathbb{R}^{q'}$, which is the solution to q estimating equations ($q' \leqslant q$),

$$\sum_{i \in \mathcal{U}} \mathbf{f}_i(\boldsymbol{\varphi}) = \mathbf{0}_q, \tag{6}$$

where $\mathbf{f}_i(\boldsymbol{\varphi}) := \mathbf{f}(\boldsymbol{\varphi}, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i, \boldsymbol{y}_i) \in \mathbb{R}^q$, $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$ and $\boldsymbol{\Phi}$ is compact. We assume that $\mathbf{f}_i(\boldsymbol{\varphi})$ is not subject to missingness. Thus, $\mathbf{f}_i(\boldsymbol{\varphi})$ is a function of the components of $\boldsymbol{y}_i$ which are not subject to missing values. In (6), $\mathbf{f}_i(\boldsymbol{\varphi})$ is a unit level function. When $\boldsymbol{\varphi}_0$ describe cluster level characteristics, we use $\mathbf{f}_i(\boldsymbol{\varphi}) = \widetilde{\mathbf{f}}_k(\boldsymbol{\varphi})\pi_{i|k}\,\nu_k^{-1}$, where $i \in \widetilde{U}_k$ and $\widetilde{\mathbf{f}}_k(\boldsymbol{\varphi})$ is a cluster level function. In this case, (6) reduces to $\sum_{k=1}^{N} \widetilde{\mathbf{f}}_k(\boldsymbol{\varphi}) = \mathbf{0}_q$, because $\sum_{i \in \widetilde{U}_k} \pi_{i|k} = \nu_k$.

The $\mathbf{f}_i(\boldsymbol{\varphi}_0)$ are called '*auxiliary variables*' in the survey sampling literature (e.g. Hartley and Rao 1968; Deville and Särndal 1992). In what follows, we shall replace $\mathbf{f}_i(\boldsymbol{\varphi}_0)$ by $\mathbf{f}_i$, because $\boldsymbol{\varphi}_0$ is a vector of known constants.

5 Empirical likelihood approach

We have two unknown parameters: the parameter of interest $\boldsymbol{\tau}_0$ and the response parameter $\boldsymbol{\lambda}_0$. Let $\boldsymbol{\psi}_0 = (\boldsymbol{\tau}_0^\top, \boldsymbol{\lambda}_0^\top)^\top$ denotes the overall parameter and $\boldsymbol{\psi} = (\boldsymbol{\tau}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\psi}_0, \boldsymbol{\psi} \in \boldsymbol{\Psi} \subset \mathbb{R}^{p+d_\xi}$ and $\boldsymbol{\Psi}$ denotes the compact parameter space of $\boldsymbol{\psi}_0$.

Consider the "*cluster-level empirical likelihood function*":

$$\ell_{\max}(\boldsymbol{\psi}) := \max_{p_k : k \in \widetilde{\boldsymbol{S}}} \left\{ \sum_{k \in \widetilde{\boldsymbol{S}}} \log p_k : \; p_k > 0, \; n \sum_{k \in \widetilde{\boldsymbol{S}}} \frac{p_k}{\pi_k} \widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}, \boldsymbol{r}) = \boldsymbol{C}^\star \right\}, \quad (7)$$

where $\boldsymbol{r}$ is defined within (1) and

$$\widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}, \boldsymbol{r}) := \left\{ \widehat{\boldsymbol{a}}_k(\boldsymbol{\psi}, \boldsymbol{r})^\top, \widehat{\boldsymbol{c}}_k^\top \right\}^\top, \qquad \widehat{\boldsymbol{c}}_k := \left\{ N n^{-1} \boldsymbol{z}_k^\top, \widehat{\mathbf{f}}_k^\top \right\}^\top, \quad (8)$$

$$\boldsymbol{C}^\star := \left(\mathbf{0}_{d_\xi + p}^\top, \boldsymbol{C}^\top \right)^\top, \qquad \boldsymbol{C} := (N n^{-1} \boldsymbol{n}_H^\top, \mathbf{0}_q^\top)^\top, \quad (9)$$

$$\widehat{\boldsymbol{a}}_k(\boldsymbol{\psi}, \boldsymbol{r}) := \sum_{i \in \boldsymbol{S}_k} \pi_{i|k}^{-1} \boldsymbol{a}_i(\boldsymbol{\psi}, \boldsymbol{r}), \quad (10)$$

$$\widehat{\mathbf{f}}_k := \sum_{i \in \boldsymbol{S}_k} \pi_{i|k}^{-1} \mathbf{f}_i, \quad (11)$$

$$\boldsymbol{a}_i(\boldsymbol{\psi}, \boldsymbol{r}) := \left[P_i(\boldsymbol{\lambda})^{-1} r_i \, \boldsymbol{g}_i(\boldsymbol{\tau})^\top, \boldsymbol{\xi}_i^\top \{ r_i - P_i(\boldsymbol{\lambda}) \} \right]^\top, \quad (12)$$

$$\boldsymbol{z}_k := (z_{k1}, \ldots, z_{kh}, \ldots, z_{kH})^\top,$$

$$\boldsymbol{n}_H := (n_1, \ldots, n_H)^\top,$$

$$z_{kh} := \begin{cases} \pi_k & \text{for } k \in U_h, \\ 0 & \text{otherwise,} \end{cases}$$

$d_\xi = \dim\{\boldsymbol{\xi}_i\}$ and log denotes the natural logarithm. The $P_i(\boldsymbol{\lambda})$ are defined by (3). Note that $\boldsymbol{g}_i(\boldsymbol{\tau})$ is function of $\boldsymbol{y}_i$. Thus, when $r_i = 0$, we have that $\boldsymbol{g}_i(\boldsymbol{\tau})$ is missing, and $r_i \boldsymbol{g}_i(\boldsymbol{\tau}) = \mathbf{0}_p$. When $\pi_{i|k}$ are unknown, survey weights can used within (10) and (11) instead of $\pi_{i|k}^{-1}$.

The $\boldsymbol{z}_k$ are the '*stratification variables*'. The $\widehat{\mathbf{f}}_k$ are related to constraint imposed by (6). The information about the parameter is included within $\widehat{\boldsymbol{a}}_k(\boldsymbol{\psi}, \boldsymbol{r})$. Expression (7) is a cluster-level function because of the sum over $k \in \widetilde{\boldsymbol{S}}$ within (7). The key idea of the paper is to show that (7) can be used for consistent point estimation and gives a pivotal empirical log-likelihood ratio statistic. One of the unique feature of the approach is the inclusion of a set of stratification constraints, $n \sum_{k \in \widetilde{\boldsymbol{S}}} p_k \pi_k^{-1} \boldsymbol{z}_k = \boldsymbol{n}_H$, given by (8) and (9), not motivated by moment conditions. The other feature is the weights π_k^{-1} included within the constraint of (7).

It can be shown that the constraint within (7) implies $\sum_{k \in \widetilde{\boldsymbol{S}}} p_k = 1$, which is known as the leading constraint. The definition (7) resembles the standard empirical likelihood function (Owen 1988), apart from the weight π_k^{-1} within the constraint. The p_k play the same role as the g-weights as in (e.g. Särndal et al 1992, p.232) or calibration factor (Deville and Särndal 1992).

Using Lagrangian multiplier, we have that

$$\ell_{\max}(\boldsymbol{\psi}) = \sum_{k\in\widetilde{\boldsymbol{S}}} \log \widehat{p}_k(\boldsymbol{\psi})\cdot$$

where

$$\widehat{p}_k(\boldsymbol{\psi}) = n^{-1}\Big\{1 + \boldsymbol{\eta}(\boldsymbol{\psi})^\top \widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}, \boldsymbol{r})\pi_k^{-1}\Big\}^{-1}\cdot$$

Here, $\boldsymbol{\eta}(\boldsymbol{\psi})$ is such that $\widehat{p}_k(\boldsymbol{\psi}) > 0$ and the following constraint holds.

$$n \sum_{k\in\widetilde{\boldsymbol{S}}} \pi_k^{-1}\widehat{p}_k(\boldsymbol{\psi})\, \widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}, \boldsymbol{r}) = \boldsymbol{C}^\star\cdot$$

We assume that $\boldsymbol{\psi}$ is such that $\boldsymbol{C}^\star$ is an inner point of the convex conical hull of $\{\widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}, \boldsymbol{r}) : k \in \widetilde{\boldsymbol{S}}\}$, so that a unique solution $\boldsymbol{\eta}(\boldsymbol{\psi})$ exists.

5.1 Point estimation

The '*maximum empirical likelihood estimator*' is

$$\widehat{\boldsymbol{\psi}} := \arg\max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \ell_{\max}(\boldsymbol{\psi})\cdot \tag{13}$$

It can be shown that $\widehat{\boldsymbol{\psi}}$ is also the solution to

$$\widehat{\boldsymbol{A}}(\boldsymbol{\psi}) := n \sum_{k\in\widetilde{\boldsymbol{S}}} \widehat{p}_k\, \pi_k^{-1}\, \widehat{\boldsymbol{a}}_k(\boldsymbol{\psi}, \boldsymbol{r}) = \boldsymbol{0}_{t+p}, \tag{14}$$

where

$$\widehat{p}_k := n^{-1}\big(1 + \boldsymbol{\eta}^\top \widehat{\boldsymbol{c}}_k\, \pi_k^{-1}\big)^{-1}\cdot \tag{15}$$

Here, $\boldsymbol{\eta}$ is such that $\widehat{p}_k > 0$ and $n\sum_{k\in\widetilde{\boldsymbol{S}}} \widehat{p}_k\, \pi_k^{-1}\, \widehat{\boldsymbol{c}}_k = \boldsymbol{C}$, where $\widehat{\boldsymbol{c}}_k$ and $\boldsymbol{C}$ are respectively defined by (8) and (9). Note that the constraint always implies $\sum_{k\in\widetilde{\boldsymbol{S}}} \widehat{p}_k = 1$.

By using (12) and (10), we have that (14) reduces to sample-based estimating equations:

$$\sum_{k\in\widetilde{\boldsymbol{S}}} \sum_{i\in\boldsymbol{S}_k} w_{i|k}\boldsymbol{\xi}_i\{r_i - P_i(\boldsymbol{\lambda})\} = \boldsymbol{0}_{d_\xi}, \tag{16}$$

$$\sum_{k\in\widetilde{\boldsymbol{S}}} \sum_{i\in\boldsymbol{S}_k} w_{i|k}\mathbf{f}_i = \boldsymbol{0}_q, \tag{17}$$

$$\sum_{k\in\widetilde{\boldsymbol{S}}} \sum_{i\in\boldsymbol{S}_k} r_i\, P_i(\boldsymbol{\lambda})^{-1} w_{i|k}\, \boldsymbol{g}_i(\boldsymbol{\tau}) = \boldsymbol{0}_p, \tag{18}$$

where $w_{i|k} := \widehat{m}_k\, \pi_{i|k}^{-1}$ and $\widehat{m}_k := n\, \widehat{p}_k\, \pi_k^{-1}$. The quantities $\widehat{m}_k$ and $\pi_{i|k}^{-1}$ are respectively the cluster-level empirical likelihood weights and the unit-level weight. The quantities $P_i(\boldsymbol{\lambda})^{-1}$ are '*propensity-score adjustments*'. In § 6, we will see that they ensure consistency of $\widehat{\boldsymbol{\psi}}$. The non-response parameter $\boldsymbol{\lambda}_0$ is estimated from (16), which is the sample-level weighted estimating equation of a generalised linear model. Equation (17) takes into account of the side information. The parameter $\boldsymbol{\tau}_0$ is estimated from the equation (18). Note that (18) (17) are sum over the non-missing unit i, with $r_i = 1$.

In the particular case when we have a simple random sample with no stratification, clustering and side information, the estimating equations (16)–(18) are indeed those obtained by Qin et al (2009), under an i.i.d. setting.

The constant $\boldsymbol{\varphi}_0$ can be estimated by including $\mathbf{f}_i(\boldsymbol{\varphi})$ within $\boldsymbol{g}_i(\boldsymbol{\tau})$, where $\mathbf{f}_i(\boldsymbol{\varphi})$ is defined in (6). Equation (17) implies that the maximum empirical likelihood estimator of $\boldsymbol{\varphi}_0$ is an almost surely constant random variable taking a single value $\boldsymbol{\varphi}_0$. This property is known as the '*calibration*' in survey sampling literature (Deville and Särndal 1992). Calibration is the consequence of the maximisation (13). In survey sampling literature, calibration is viewed as weighting procedure, rather than the consequence of the maximisation of an empirical likelihood function.

5.2 Profile empirical likelihood ratio statistic

To allow hypotheses testing and computation of confidence intervals, we need a pivotal statistics. We propose to use the *profile empirical log-likelihood ratio statistic* defined by (19). Theorem 1 shows that (19) is a pivotal.

Consider that the parameter of interest $\boldsymbol{\theta}_0$ is a sub-vector of $\boldsymbol{\psi}_0 = (\boldsymbol{\tau}_0^\top, \boldsymbol{\lambda}_0^\top)^\top$. The remaining parameter is $\boldsymbol{\mu}_0$; that is, $\boldsymbol{\psi}_0 = (\boldsymbol{\theta}_0^\top, \boldsymbol{\mu}_0^\top)^\top$. The respective maximum empirical likelihood estimators are denoted $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\mu}}$. In practice, $\boldsymbol{\lambda}_0$ is usually not a parameter of interest and is therefore part of $\boldsymbol{\mu}_0$.

Let $\ell_{\max}(\boldsymbol{\theta}, \boldsymbol{\mu}) := \ell_{\max}(\boldsymbol{\psi})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\mu} \in \boldsymbol{\mathcal{M}}$ and $\boldsymbol{\psi} = (\boldsymbol{\tau}^\top, \boldsymbol{\lambda}^\top)^\top = (\boldsymbol{\theta}^\top, \boldsymbol{\mu}^\top)^\top$. Here, $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{M}}$ denote the parameter space of $\boldsymbol{\theta}_0$ and $\boldsymbol{\mu}_0$. The *profile empirical log-likelihood ratio statistic* is defined by the following function of $\boldsymbol{\theta}$.

$$\widehat{R}(\boldsymbol{\theta}) := 2\big\{\ell_{\max}(\widehat{\boldsymbol{\psi}}) - \max_{\boldsymbol{\mu}\in\boldsymbol{\mathcal{M}}} \ell_{\max}(\boldsymbol{\theta}, \boldsymbol{\mu})\big\}\cdot \tag{19}$$

It can be shown that $\ell_{\max}(\widehat{\boldsymbol{\psi}}) = \sum_{k\in\widetilde{\boldsymbol{S}}} \log(\widehat{p}_k)$, where $\widehat{p}_k$ is defined by (15). Theorem 1 in §6.2 shows that (19) is pivotal. Thus, (19) can used as traditional ratio statistic to test or construct confidence regions for $\boldsymbol{\theta}_0$ (see §8). The algorithm proposed by Oğuz-Alper and Berger (2016, pp 457–458) can be used to compute (19).

6 Asymptotic results

The asymptotic framework considered is based on an infinite nested sequence of sampling designs, samples and populations as in Isaki and Fuller (1982). We assume that $n \to \infty$, where n is the number of clusters sampled. The number of strata H is constant. Let $o_p(\cdot)$ and $O_p(\cdot)$ be the orders of convergence in probability with respect to the response mechanism and the sampling design. The orders $\boldsymbol{\mathcal{O}}_p(a)$ and $\boldsymbol{o}_p(a)$ are matrices (or vectors) which are such that $\|\boldsymbol{\mathcal{O}}_p(a)\| = O_p(a)$ and $\|\boldsymbol{o}_p(a)\| = o_p(a)$, where $\|\cdot\|$ denotes the Frobenius norm.

6.1 Regularity conditions

We assume the following conditions

[C1] $\max_{k=1,\ldots,N} \widetilde{N}_k = O(1)\cdot$

[C2] $\max_{i \in U}(\rho_i^{-1}) = O(1)\cdot$

[C3] $Nn^{-1} \max_{k=1,\ldots,N}(\pi_k) = O(1)\cdot$

[C4] $N^{-1}n \max_{k=1,\ldots,N}(\pi_k^{-1}) = O(1)\cdot$

[C5] $\max_{k=1,\ldots,N} \max_{i \in \widetilde{U}_k}(\pi_{i|k}^{-1}) = O(1)$, where $\pi_{i|k}$ is the conditional inclusion probability of a unit i in $\widetilde{U}_k$.

[C6] $n_h n^{-1} = \psi_h$, where ψ_h is a strictly positive fixed constant that does not vary as $n \to \infty$ ($\forall h = 1, \ldots, H$).

[C7] $N_h N^{-1} = \Psi_h$, where Ψ_h is a strictly positive fixed constant that does not vary as $n \to \infty$ ($\forall h = 1, \ldots, H$).

[C8] There exists a set of vectors of constants $\bar{\boldsymbol{\mathcal{C}}}_h$, such that

$$\sum_{h=1}^{H} N_h \bar{\boldsymbol{\mathcal{C}}}_h = \boldsymbol{C}^\star, \quad n_h^{\frac{1}{2}}\left(N_h^{-1}\widehat{\boldsymbol{C}}_{0h} - \bar{\boldsymbol{\mathcal{C}}}_h\right) = \boldsymbol{\mathcal{O}}_p(1) \quad \text{and} \quad \bar{\boldsymbol{\mathcal{C}}}_h = \boldsymbol{\mathcal{O}}_p(1),$$

$\forall\ h = 1, \ldots H$, where

$$\widehat{\boldsymbol{C}}_{0h} := \sum_{k \in \widetilde{\boldsymbol{S}}_h} \pi_k^{-1} \widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}_0, \boldsymbol{r})\cdot \tag{20}$$

[C9] $n^{-\frac{1}{2}} \max_{k \in \widetilde{\boldsymbol{S}}} \|\widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}_0, \boldsymbol{r})\| = o_p(1)\cdot$

[C10] $N^{-\mu} n^{\mu-1} \sum_{k \in \widetilde{\boldsymbol{S}}} \pi_k^{-\mu} \|\widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}_0, \boldsymbol{r})\|^\mu = O_p(1), \qquad (\mu = 1,\ 2,\ 3,\ 4)\cdot$

[C11] There exists a matrix of negative constants $\boldsymbol{S}$ such that $\widehat{\boldsymbol{S}}_0 - \boldsymbol{S} = \boldsymbol{o}_p(1)$, $\boldsymbol{S} = \boldsymbol{\mathcal{O}}(1)$ and $-\boldsymbol{S}$ is positive definite, where

$$\widehat{\boldsymbol{S}}_0 := -nN^{-2} \sum_{k \in \widetilde{\boldsymbol{S}}} \pi_k^{-2} \widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}_0, \boldsymbol{r})\, \widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}_0, \boldsymbol{r})^\top\cdot \tag{21}$$

[C12] $N^{-1} \sum_{i \in \mathcal{U}} \|\boldsymbol{\xi}_i\|^2 = O(1)\cdot$

[C13] $N^{-1}\sum_{i\in\mathcal{U}}\|\boldsymbol{g}_i(\boldsymbol{\tau}_0)\|^2 = O(1)\cdot$

[C14] $N^{-1}\sum_{i\in\mathcal{U}}\|\mathbf{f}_i\|^2 = O(1)\cdot$

[C15] $N^{-1}\dfrac{\partial\widehat{\boldsymbol{A}}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}$ is continuous in $\boldsymbol{\psi}\in\boldsymbol{\Psi}$, with $\left\|N^{-1}\dfrac{\partial\widehat{\boldsymbol{A}}(\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}_0}\right\| \asymp_p 1\cdot$

[C16] $N^{-1}\dfrac{\partial^2\widehat{\boldsymbol{A}}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^2} = \boldsymbol{\mathcal{O}}_p(1)$ uniformly for $\boldsymbol{\psi}\in\boldsymbol{\Psi}\cdot$

[C17] $\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 = \boldsymbol{o}_p(1)\cdot$

[C18] $N^{-1}\ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0) \xrightarrow{d} \boldsymbol{N}(\mathbf{0},\boldsymbol{V}_0)$,

where $\boldsymbol{I}$ denotes the identity matrix,

$$\ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0) := \widehat{\boldsymbol{A}}_{0\pi} + \widehat{\boldsymbol{B}}_0^{\top}(\boldsymbol{C}-\widehat{\boldsymbol{C}}), \tag{22}$$

$$\widehat{\boldsymbol{A}}_{0\pi} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-1}\widehat{\boldsymbol{a}}_{0k}, \tag{23}$$

$$\widehat{\boldsymbol{B}}_0 := \Big(\sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\,\widehat{\boldsymbol{c}}_k\,\widehat{\boldsymbol{c}}_k^{\top}\Big)^{-1}\sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\,\widehat{\boldsymbol{c}}_k\,\widehat{\boldsymbol{a}}_{0k}^{\top}, \tag{24}$$

$$\widehat{\boldsymbol{C}} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-1}\widehat{\boldsymbol{c}}_k,$$

$$\boldsymbol{V}_0 := \mathbb{V}\big\{N^{-1}\ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0)\big\} \tag{25}$$

and $\widehat{\boldsymbol{a}}_{0k} := \widehat{\boldsymbol{a}}_k(\boldsymbol{\psi}_0,\boldsymbol{r})$. The operator $\mathbb{V}(\cdot)$ denotes the variance with respect to the response mechanism and the sampling design.

We assume that there exist positive random variables $\mathcal{H}_i$, $\mathcal{F}_i$ and $\mathcal{B}_{ij}$ which do not depends on n and N, such that for all n

[C19] $\mathbb{E}(\mathcal{H}_i) < \infty$ and $nN^{-2}\sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\,\widehat{a}_{0ki}^2 \leqslant \mathcal{H}_i, \quad \forall i,$

[C20] $\mathbb{E}(\mathcal{F}_i) < \infty$ and $nN^{-2}\sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\,\widehat{\mathrm{f}}_{0ki}^2 \leqslant \mathcal{F}_i, \quad \forall i$

[C21] $\mathbb{E}(\mathcal{B}_{ij}) < \infty$ and $|\widehat{B}_{0ij}| \leqslant \mathcal{B}_{ij}, \quad \forall i,j;$

where $\widehat{\mathrm{f}}_{0ki}$ and $\widehat{a}_{0ki}$ are respectively the i-th component of $\widehat{\mathbf{f}}_{0k} := \widehat{\mathbf{f}}_k$ and $\widehat{\boldsymbol{a}}_{0k}$. Here, $\widehat{B}_{0ij}$ is the (i,j) component of $\widehat{\boldsymbol{B}}_0$ defined by (24). The operator $\mathbb{E}(\cdot)$ denotes the expectation with respect to the response mechanism and the sampling design.

Condition [C1] ensure that the clusters' sizes are bounded. This condition is usually met in practice, because these sizes are rarely large. Condition [C2] excludes situations when the response propensities tend to zero. Conditions [C4] and [C3] are standard requirement for π_k (e.g. Krewski and Rao 1981; Fuller 2009, p.49). It excludes π_k disproportionally smaller or larger than n/N. Condition [C5] means that the $\pi_{i|k}$ are not disproportionally small. Conditions [C6] and [C7] imply that n_h (and N_h) tends to ∞, with the same rate as n

(and N). The condition [C8] assumes that the law of large numbers holds for (20). Using Markov's inequality it can be shown that [C9] holds when $\mathbb{E}(\|\widehat{\boldsymbol{c}}_k^{\star}(\boldsymbol{\psi}_0,\boldsymbol{r})\|^4) = O(1)$ (Chen and Sitter 1999, Appendix 2). The condition [C10] is a Lyapounov's type conditions for the existence of sample moments (Krewski and Rao 1981§6.4.1). Condition [C11] assumes that the matrix of moments (21) is consistent. Condition [C12], [C13] and [C14] are standard moment conditions. Conditions [C15] and [C16] are smoothness requirement for $\widehat{\boldsymbol{A}}(\boldsymbol{\psi})$ (e.g Godambe and Thompson 1974). Both ensure that the Taylor expansion of $\widehat{\boldsymbol{A}}(\boldsymbol{\psi})$ exists. Condition [C17] is a standard requirement for solutions to estimating equations (e.g. Godambe and Thompson 2009, p 90). It relies on conditions on $\boldsymbol{A}(\boldsymbol{\psi})$ and $\widehat{\boldsymbol{A}}(\boldsymbol{\psi})$ proposed by Van Der Vaart (1998 §5). Condition [C18] assumes that the central limit theorem holds for $N^{-1}\ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0)$. It can be justified by Fuller's (2009, Ch.2) regularity conditions. Conditions [C19]–[C21] ensure that an estimator of (25) is asymptotically unbiased (see Lemma 2).

6.2 Pivotal property of (19) and $\sqrt{n}$-consistency

The pivotal property is based on the consistency of the following variance.

$$\widehat{\boldsymbol{V}}_0 := N^{-2}\sum_{h=1}^{H}\Big(\sum_{k\in\widetilde{\boldsymbol{S}}_h}\pi_k^{-2}\widehat{\boldsymbol{e}}_k\widehat{\boldsymbol{e}}_k^{\top} - \frac{1}{n_h}\sum_{k\in\widetilde{\boldsymbol{S}}_h}\pi_k^{-1}\widehat{\boldsymbol{e}}_k\sum_{\ell\in\widetilde{\boldsymbol{S}}_h}\pi_\ell^{-1}\widehat{\boldsymbol{e}}_\ell^{\top}\Big), \tag{26}$$

where

$$\begin{aligned}\widehat{\boldsymbol{e}}_k &:= \widehat{\boldsymbol{a}}_{0k} - \widehat{\boldsymbol{b}}_0^{\top}\widehat{\mathbf{f}}_{0k},\\ \widehat{\boldsymbol{b}}_0 &:= \big(\widehat{\boldsymbol{S}}_{\mathrm{ff}} - \widehat{\boldsymbol{S}}_{z\mathrm{f}}^{\top}\widehat{\boldsymbol{S}}_{zz}^{-1}\widehat{\boldsymbol{S}}_{z\mathrm{f}}\big)^{-1}\big(\widehat{\boldsymbol{S}}_{\mathrm{f}a} - \widehat{\boldsymbol{S}}_{z\mathrm{f}}^{\top}\widehat{\boldsymbol{S}}_{zz}^{-1}\widehat{\boldsymbol{S}}_{za}\big),\end{aligned} \tag{27}$$

$$\widehat{\boldsymbol{S}}_{zz} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\boldsymbol{z}_k\boldsymbol{z}_k^{\top}, \qquad \widehat{\boldsymbol{S}}_{z\mathrm{f}} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\boldsymbol{z}_k\widehat{\mathbf{f}}_{0k}^{\top}, \qquad \widehat{\boldsymbol{S}}_{\mathrm{ff}} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\widehat{\mathbf{f}}_{0k}\widehat{\mathbf{f}}_{0k}^{\top},$$

$$\widehat{\boldsymbol{S}}_{\mathrm{f}a} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\widehat{\mathbf{f}}_{0k}\widehat{\boldsymbol{a}}_{0k}^{\top}, \qquad \widehat{\boldsymbol{S}}_{za} := \sum_{k\in\widetilde{\boldsymbol{S}}}\pi_k^{-2}\boldsymbol{z}_k\widehat{\boldsymbol{a}}_{0k}^{\top},$$

$\widehat{\boldsymbol{a}}_{0k} := \widehat{\boldsymbol{a}}_k(\boldsymbol{\psi}_0,\boldsymbol{r})$ and $\widehat{\mathbf{f}}_{0k} := \widehat{\mathbf{f}}_k$.

Lemma 1 *There exists a vector* $\boldsymbol{\beta}_0 = \mathcal{O}(1)$ *such that* $\widehat{\boldsymbol{b}}_0 - \boldsymbol{\beta}_0 = o_p(1)$, *where* $\widehat{\boldsymbol{b}}_0$ *is defined by* (27).

Lemma 1 follows from [C11] and a first-order Taylor expansion of $\widehat{\boldsymbol{b}}_0$.

Consider

$$\widetilde{\boldsymbol{V}}_0 := N^{-2}\sum_{h=1}^{H}\Big(\sum_{k\in\widetilde{\boldsymbol{S}}_h}\pi_k^{-2}\widehat{\boldsymbol{\epsilon}}_k\widehat{\boldsymbol{\epsilon}}_k^{\top} - \frac{1}{n_h}\sum_{k\in\widetilde{\boldsymbol{S}}_h}\pi_k^{-1}\widehat{\boldsymbol{\epsilon}}_k\sum_{\ell\in\widetilde{\boldsymbol{S}}_h}\pi_\ell^{-1}\widehat{\boldsymbol{\epsilon}}_\ell^{\top}\Big), \tag{28}$$

where

$$\widehat{\boldsymbol{\epsilon}}_k := \widehat{\boldsymbol{a}}_{0k} - \boldsymbol{\beta}_0^\top \widehat{\mathbf{f}}_{0k}, \tag{29}$$

and $\boldsymbol{\beta}_0$ is defined in Lemma 1. Note that $\widetilde{\boldsymbol{V}}_0$ is a function of the constant $\boldsymbol{\beta}_0$. On the other hand, $\widehat{\boldsymbol{V}}_0$ is a function of the random variable $\widehat{\boldsymbol{b}}_0$.

Lemma 2 *When $nN^{-1} = o(1)$, we have that*

$$n\boldsymbol{V}_0 = \boldsymbol{\mathcal{O}}(1), \tag{30}$$

$$n\mathbb{E}(\widetilde{\boldsymbol{V}}_0) = n\boldsymbol{V}_0 + \boldsymbol{o}(1), \tag{31}$$

where $\boldsymbol{V}_0$ is defined by (25).

The proof can be found in Appendix C of the online supplement.

Lemma 3 *When $nN^{-1} = o(1)$, we have that*

$$n(\widehat{\boldsymbol{V}}_0 - \boldsymbol{V}_0) = \boldsymbol{o}_p(1), \tag{32}$$

where $\widehat{\boldsymbol{V}}_0$ is defined by (26).

The proof can be found in Appendix B in the online supplement.

Theorem 1 *Assuming that $\widehat{\boldsymbol{c}}_k^\star(\boldsymbol{\psi}, \boldsymbol{r})$ is differentiable with respect to $\boldsymbol{\mu}$ (see* §5.2)*, conditions* [C1]–[C21] *imply*

$$\widehat{R}(\boldsymbol{\theta}_0) = N^{-2}\ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0)^\top (\boldsymbol{I} - \widehat{\boldsymbol{A}}_0)\ \widehat{\boldsymbol{V}}_0^{-1}\ \ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0) + n^{-\frac{1}{2}} O_p(1), \tag{33}$$

where,

$$\widehat{\boldsymbol{A}}_0 := \widehat{\boldsymbol{V}}_0^{-\frac{1}{2}} \widehat{\boldsymbol{\nabla}}_0 \big(\widehat{\boldsymbol{\nabla}}_0^\top \widehat{\boldsymbol{V}}_0^{-1} \widehat{\boldsymbol{\nabla}}_0\big)^{-1} \widehat{\boldsymbol{\nabla}}_0^\top \widehat{\boldsymbol{V}}_0^{-\frac{1}{2}}, \tag{34}$$

$$\widehat{\boldsymbol{\nabla}}_0 := N^{-1} \frac{\partial \ddot{\boldsymbol{A}}(\boldsymbol{\psi})}{\partial \boldsymbol{\mu}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}.$$

The proof can be found in Appendix B in the online supplement.

Using [C18] and Lemma 3, the Slutsky's Theorem implies

$$\widehat{\boldsymbol{V}}_0^{-\frac{1}{2}}\ N^{-1}\ddot{\boldsymbol{A}}(\boldsymbol{\psi}_0) \xrightarrow{d} \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{I})\cdot \tag{35}$$

Thus, when $nN^{-1} = o(1)$, expressions (35) and (33) imply

$$\widehat{R}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_{df=t}, \tag{36}$$

because $(\boldsymbol{I} - \widehat{\boldsymbol{A}}_0)$ is a symmetric idempotent matrix with trace t, where $t = \dim(\boldsymbol{\theta}_0)$.

Theorem 2 *Under* [C3], [C4], [C8], [C9], [C10], [C11], [C15], [C16] *and* [C17], *we have that $\widehat{\boldsymbol{\psi}}$ is $\sqrt{n}$-consistent; that is,*

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) = \boldsymbol{\mathcal{O}}_p(1)\cdot$$

The proof of Theorems 1 and 2 can be found in Appendix B of the online supplement.

7 Cluster sampling with large sampling fractions

In this §, we extend the approach proposed when the clusters are selected without replacement and with a large sampling fraction n/N. For instance, this is can be the case for the "*National Health and Nutrition Examination Survey*" (National Center for Health Statistics 2016). Note that in §5, we allow the within-cluster sampling fractions $\nu_k/\widetilde{N}_k$ to be large. For the extreme case when $\nu_k/\widetilde{N}_k = 1$, we have a single stage design and Berger and Torres's (2016) empirical likelihood approach for single stage designs with large sampling fractions can used. This §'s extension is not be based upon Berger and Torres's (2016) approach.

Our point estimator $\widehat{\boldsymbol{\psi}}$ is still $\sqrt{n}$-consistent when n/N is large, because Theorem 2 does not rely on $n/N = o(1)$. It is also not necessary to have $n/N = o(1)$, for Theorem 1 to hold. However, (36) may not hold any longer when n/N is large; because of the condition of Lemma 3 is not satisfied. that is, $\widehat{R}(\boldsymbol{\theta}_0)$ converges to a distribution which is different from a χ^2-distribution. Since (33) holds when n/N is large, standard results on the distribution of quadratic forms (e.g. Scheffé, 1959 p.418; Rao, 1973 and Wu et al, 2017) can be used to show that (33) and [C18] imply that $\widehat{R}(\boldsymbol{\theta}_0)$ converges to a linear combination of χ^2-distribution; that is,

$$\widehat{R}(\boldsymbol{\theta}_0) \xrightarrow{d} \sum_{\ell=1}^{p} \lambda_\ell\, \mathcal{Z}_\ell^2, \tag{37}$$

where $\mathcal{Z}_1, \ldots, \mathcal{Z}_p$ are independently distributed standard normal variables and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of

$$\boldsymbol{\mathcal{L}}_0 = \boldsymbol{V}_0(\boldsymbol{I} - \widehat{\boldsymbol{A}}_0)\ \widehat{\boldsymbol{V}}_0^{-1} \tag{38}$$

where $\boldsymbol{V}_0$, $\widehat{\boldsymbol{V}}_0$ and $\widehat{\boldsymbol{A}}_0$ are respectively given by (25), (26) and (34). In Appendix A, we propose an estimator $\widehat{\boldsymbol{\mathcal{L}}}_0$ of (38). Let $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\boldsymbol{\mathcal{L}}}_0$. The inference can be based upon $\widehat{R}(\boldsymbol{\theta})$ and the quantiles of $\sum_{\ell=1}^{p} \lambda_\ell\, \mathcal{Z}_\ell^2$ estimated numerically from the empirical distribution of $\sum_{\ell=1}^{p} \widehat{\lambda}_\ell\, \mathcal{Z}_\ell^2$.

When $n/N = o(1)$, Lemma 3 holds and by substituting $\boldsymbol{V}_0$ by $\widehat{\boldsymbol{V}}_0$ within (38), we obtain t eigenvalues equal to 1 and $p-t$ eigenvalues equal to 0, which indeed implies (36).

It is common practise to have $\boldsymbol{\theta}_0$ scalar ($t = 1$), for example, when we are interested in confidence intervals of a component of $\boldsymbol{\psi}_0$. In this case, we have a single strictly positive eigenvalue, say λ_1 and $p-1$ eigenvalues equal to 0. Hence, (37) reduces to

$$\widehat{R}(\boldsymbol{\theta}_0)\lambda_1^{-1} \xrightarrow{d} \chi^2_{df=1} \quad \text{when } t = 1\cdot \tag{39}$$

After replacing λ_1 by its estimates, (39) can be used for inference.

The simulation study on §8.4 shows that confidence intervals based on (39) are slightly shorter than those based (36). The more conservative confidence interval based on (36) seems preferable. We have a simple interpretation for the minor differences between the confidence intervals based on (39) and (36). The variance $\widehat{\boldsymbol{V}}_0$ within the quadratic form (33) converges to the random matrix $\widetilde{\boldsymbol{V}}_0$ (see (B.27) in Appendix B in the online supplement). The expectation of $\widetilde{\boldsymbol{V}}_0$ is a sum of a between and within-cluster variances (see (C.15) and (C.37) in the online supplement). Thus, $\widehat{\boldsymbol{V}}_0$ captures both terms of the two-stage variance. However, the between variance does not contain any finite population corrections, such as joint-inclusion probabilities or Hájek's (1964) finite population correction. Hence, the lack of finite population corrections increases the bias of $\widehat{\boldsymbol{V}}_0$. One of the terms due to non-response is the variance due to the non-response mechanism of the design expectation. This non-response term is of order N^{-1} (see (C.13) in the online supplement), which is still small compared to the overall variance of order n^{-1} (see (30)). When $n/N \to 0$, this term is ignored within $\widehat{\boldsymbol{V}}_0$, because it is asymptotically negligible. When $n/N \not\to 0$, the absence of this non-response term within $\widehat{\boldsymbol{V}}_0$ decreases the bias of $\widehat{\boldsymbol{V}}_0$. Finally, the increase in bias due to the lack of finite population correction and the decreases in bias due to the absence of one of the non-response term may compensate each other, and produce a negligible bias for $\widehat{\boldsymbol{V}}_0$, even when $n/N \not\to 0$.

8 Numerical results

We consider the following parameters of interest: population mean and poisson regression parameters, multiple logistic regression parameters, quantiles and distribution functions. We consider that $\boldsymbol{\theta}_0$ is a scalar θ_0. Thus, the 95% confidence intervals can be constructed using (36); that is,

$$\textsc{ci}(\theta_0) := \Big\{\theta : \widehat{R}(\theta) \leqslant 3.8415\Big\}, \tag{40}$$

where 3.8415 is the upper 95% quantile of the χ^2-distribution with one degree of freedom. In §§ 8.1, 8.2 and 8.3, we report the observed coverages of (40). Single-stage sampling without side information is considered in §§ 8.1 and 8.2. This allows to investigate the effect of non-response without the clustering effect. In §§ 8.3 and 8.4, we consider cluster (two-stage) sampling designs. Side information is considered in §8.3. In §8.4, we consider large sampling fractions.

In all the simulation studies, we have multidimensional parameters, because of the response parameter needs to be estimated and is part of $\boldsymbol{\psi}_0$. Indeed, we have $\boldsymbol{\psi}_0 = (\boldsymbol{\tau}_0^\top, \boldsymbol{\lambda}_0^\top)^\top$, where $\boldsymbol{\tau}_0$ is the parameter of interest and $\boldsymbol{\lambda}_0$ is the response parameter. The profile empirical log-likelihood ratio statistic (19) allows to construct the confidence interval (40) for each scalar components θ_0 of $\boldsymbol{\tau}_0$.

The simulation was done in R (R Development Core Team 2014). Some of the R codes used in this §, are available on the author's web-page:
`http://www.yvesberger.co.uk`.

8.1 Population mean

Consider artificial populations of $\mathcal{N} = 10\,000$ values y_i generated from a skewed distribution given by

$$y_i = 3 + \zeta_i + x_i + \psi(e_i - 1),$$

where $\zeta_i \sim \exp(1)$, $x_i \sim \exp(1)$ and $e_i \sim \chi^2_{df=1}$. Five populations are generated with $\psi = 20,\ 3.5,\ 1.5,\ 0.6$ and 0.1. We consider a single-stage randomised systematic sampling design (e.g. Hartley and Rao 1962); that is, $\boldsymbol{S}_k = \widetilde{U}_k = \{k\}$ and $\nu_k = 1$. We select $10\,000$ samples of size 100. The π_i are proportional to $|\zeta_i| + 2$. Different values of ψ allows the correlation between π_i and y_i to vary between 0.02 and 0.7. Missing values for y_i are generated from (2) with $\digamma$ in (3) being the logit function, $\boldsymbol{\xi}_i = (1, \xi_i)^\top$ and $\boldsymbol{\lambda}_0 = (-1, 1)^\top$. Here, $\xi_i \sim \Gamma(\text{shape} = 1,\ \text{scale} = 2)$. The response propensities (4) lie between 0.27 and 1, with an average of 0.6. Side information is not considered. The parameter of interest is the mean $\boldsymbol{\tau}_0 := N^{-1} \sum_{i \in \mathcal{U}} y_i$. We use $\boldsymbol{g}_i(\boldsymbol{\tau}) = y_i - \theta$ with $\boldsymbol{\tau} = \theta$.

The customary two-phase point estimator of $\boldsymbol{\tau}_0$ is

$$\widehat{\theta}_c := \widehat{N}^{-1} \sum_{i \in \boldsymbol{S}} r_i (\pi_i \widehat{\rho}_i)^{-1} y_i, \tag{41}$$

where $\widehat{N} := \sum_{i \in \boldsymbol{S}} r_i (\pi_i \widehat{\rho}_i)^{-1}$ is an estimator of N and $\widehat{\rho}_i$ are the fitted probabilities of a logistic model with an intercept and ξ_i. The customary two-phase variance estimator (Särndal et al 1992 §9.4) is

$$\begin{aligned}\widehat{\mathbb{V}}_2(\widehat{\theta}_c) := \widehat{N}^{-2} \Big[& \sum_{i \in \boldsymbol{S}} r_i \sum_{j \in \boldsymbol{S}} r_j \big(\pi_i^{-1} \pi_j^{-1} - \pi_{ij}^{-1}\big) \widehat{\rho}_{ij}^{-1} (y_i - \widehat{\theta}_c)(y_j - \widehat{\theta}_c) \\ & + \sum_{i \in \boldsymbol{S}} r_i \, \pi_i^{-2} \, \widehat{\rho}_i^{-2} \, (y_i - \widehat{\theta}_c)^2 (1 - \widehat{\rho}_i) \Big],\end{aligned} \tag{42}$$

where $\widehat{\rho}_{ij} = \widehat{\rho}_i \, \widehat{\rho}_j$ for $i \neq j$ and $\widehat{\rho}_{ii} = \widehat{\rho}_i$. Here, $\pi_{ij} := \mathbb{P}_d(d_i = 1, d_j = 1)$ denote joint-inclusion probabilities, which can be computed with Hartley and Rao's (1962) formula.

[Table 1]

In Table 1, we have the observed coverages of the empirical likelihood confidence interval (40) and the standard confidence interval based on (42) and the normality assumption for $\widehat{\theta}_c$. We observe a low coverage for the standard approach, because of the skewness of the data. We also have significantly large right error rates and low left error rates. The coverage of the empirical likelihood approach are closed to 95%. They are however significantly different from 95%, because 10 000 sample are selected. The left and right error rates are more balanced, with some not significantly different from 2.5%. We do not observed difference between the mean squared errors (MSE) of the empirical likelihood point estimator and (41), because side information is not used.

8.2 Poisson regression

Consider artificial populations of approximately 40 000 values y_i generated from a poisson distribution with mean $\vartheta x_i u_i$, where $x_i \sim \mathcal{N}(10, \sigma^2)$ and $u_i \sim \mathcal{N}(100, 30^2)$ is an offset. We exclude the units with $\vartheta x_i u_i \leqslant 0$. Several population data are generated with ϑ = -0.5, -0.4, -0.2, 0.005; and σ = 0.5, 1.0, 2.0, 3.0 or 4.0. The parameter of interest is the poisson regression parameter $\theta_0 = \boldsymbol{\tau}_0$, which is the solution to (5) with

$$\boldsymbol{g}_i(\boldsymbol{\tau}) = x_i \big\{ y_i - u_i \exp(\theta x_i) \big\} \cdot$$

Here, $\boldsymbol{\tau} = \theta$ and $\boldsymbol{y}_i = (y_i, x_i, u_i)^\top$. For simplicity, we did not include an intercept and side information is not considered.

We select 10 000 single-stage randomised systematic samples, as described in §8.1. The sample size is 400. The π_i are proportional to $|\zeta_i + 10|$, where $\zeta_i = 0.7(y_i - \overline{Y})\sigma_y^{-1} + e_i$, $e_i \sim \mathcal{N}(0, 0.51)$, $\overline{Y} = N^{-1} \sum_{i \in \mathcal{U}} y_i$ and $\sigma_y^2 = (N - 1)^{-1} \sum_{i \in \mathcal{U}} (y_i - \overline{Y})^2$. The correlation between π_i and y_i is approximately 0.7.

Missing values for $\boldsymbol{y}_i$ are generated from (2) with $\digamma$ in (3) being the logit function, $\boldsymbol{\xi}_i = (1, \xi_i)^\top$ and $\boldsymbol{\lambda}_0 = (-8, 1)^\top$. Here, $\xi_i = 0.8(x_i - \overline{X})\sigma_x^{-1} + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 0.36)$, $\overline{X} = N^{-1} \sum_{i \in \mathcal{U}} x_i$ and $\sigma_x^2 = (N - 1)^{-1} \sum_{i \in \mathcal{U}} (x_i - \overline{X})^2$. The correlation between ξ_i and x_i is approximately 0.8. The response propensities (3) lie between 0.02 and 1, with an average of 0.88. We have a non-MAR response mechanism, because we may have a correlation between y_i and ρ_i.

We compare the proposed empirical likelihood approach proposed with the naïve approach based on maximum likelihood from the set of non-missing values. This naïve approach is not likely to perform well, because it does not contain adjustment for missing data and y_i can be correlated with ρ_i. This naïve approach is used as a benchmark.

[Table 2]

In Table 2, the 9-th column shows that the MSE of the empirical likelihood point estimator is smaller than the MSE of the naïve estimator, when σ is small. We also notice that the overall coverage is well below 95%, with coverages increases with ϑ. The empirical likelihood approach gives coverages closer to 95%, when the correlation between y_i and ρ_i is negligible. Even with large correlation (non-MAR), we obtain acceptable coverages. The error rates are also significantly different than 2.5% with the naïve approach. With the empirical likelihood approach, the coverages and error rates are respectively not significantly different from 95% and 2.5%, except in very few cases, despite that 10 000 sample were selected.

8.3 Logistic model with income and living conditions data

The "*European Union Statistics on Income and Living Conditions*" (EU-SILC) collects information on income, living conditions and poverty (Eurostat 2012). We use Alfons et al (2011) synthetic dataset called AMELIA, based on EU-SILC.

AMELIA maintains the association between key variables. A full description of AMELIA can be found in Alfons et al (2011).

We consider a subset of AMELIA defined by the 3 regions (PROV=1, 2, 3), and individuals between 19 and 79 years of age. These three regions will be used as strata. This subset is replicated twenty times to create an artificial population of 1 539 368 individuals. The strata sizes are 148 236, 567 376 and 823 756 respectively. Individuals are grouped into communities (variable CIT), which play the role as clusters. The first stratum contains 1420 clusters, the second stratum contains 2 540 clusters, and the third stratum contains 2640 clusters. We select respectively 72, 128 and 132 clusters within strata 1, 2 and 3. A randomised systematic sample of clusters are selected with probabilities proportional to clusters' sizes. Within each clusters selected, 20% of individuals (with a minimum of 5) are sampled with simple random sampling. Missing values are generated from (2) with F being the logit function,

$$\begin{aligned}\boldsymbol{\xi}_i &= [1, \xi_i^{(a)}, \xi_i^{(b)}, \delta\{\text{Urb}_i = 1\}, \delta\{\text{Urb}_i = 3\}]^\top,\\ \boldsymbol{\lambda}_0 &= (0.5, -0.2, 0.2, -0.2, 0.2)^\top,\end{aligned}$$

where $\xi_i^{(a)}$ and $\xi_i^{(b)}$ are values generated independently form a Bernoulli(0.05) distribution. Here Urb$_i$ is the level of urbanisation of individual i: Urb$_i = 1$ for densely-populated areas, Urb$_i = 2$ for intermediate-populated areas and Urb$_i = 3$ for thinly-populated areas. The function $\delta\{A\} = 1$ when A is true and $\delta\{A\} = 0$, otherwise. The response probabilities generated lies between 0.52 and 0.72.

The model of interest is the following logistic model

$$\text{Logit}\{Pr(\text{Unemp}_i = 1)\} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Educ}_i + \beta_3 \text{Married}_i + \beta_4 \text{Male}_i,$$

where Unemp$_i = 1$ if the individual i is unemployed and Unemp$_i = 0$ otherwise. The explanatory variables are:

(I) Age$_i$, the age of i,
(II) Educ$_i = 1$ if the Highest ISCED level attained is strictly larger than 3 and Educ$_i = 0$ otherwise,
(III) Married$_i = 1$ if i is married and Married$_i = 0$ otherwise,
(IV) Male$_i = 1$ if i is male and Male$_i = 0$ otherwise.

The estimating function of this logistic model is

$$\boldsymbol{g}_i(\boldsymbol{\tau}) := \boldsymbol{x}_i y_i - \boldsymbol{x}_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\tau})\{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\tau})\}^{-1}, \tag{43}$$

where $y_i :=$ Unemp$_i$ and $\boldsymbol{x}_i := (\text{Age}_i, \text{Educ}_i, \text{Married}_i, \text{Male}_i)^\top$. The parameter is $\boldsymbol{\psi}_0 = (\boldsymbol{\tau}_0^\top, \boldsymbol{\lambda}_0^\top)^\top$, where $\boldsymbol{\tau}_0$ are the regression coefficients of $\boldsymbol{x}_i$ in the model (43). The response parameter $\boldsymbol{\lambda}_0$ contains the coefficient of $\boldsymbol{\xi}_i$ in the model (2). Hence, $\boldsymbol{\psi}_0 \in \mathbb{R}^{10}$. The side information are the fractions of individuals within densely-populated areas (35.57%), and within intermediate-populated areas (22.73%). Thus, $\mathbf{f}_i = \{\delta(\text{Urb}_i = 1), \delta(\text{Urb}_i = 2)\}^\top - \boldsymbol{\varphi}_0$, where

$\boldsymbol{\varphi}_0 = (0.3557, 0.2273)^\top$. The fraction of remaining thinly-populated areas is redundant and does not need to be included within $\boldsymbol{\varphi}_0$.

We compare the proposed empirical likelihood approach proposed with the naïve approach based on maximum likelihood from the set of non-missing values. This naïve approach is used as a benchmark. We also consider Owen's (2001) customary empirical likelihood approach (Column "*Cust.* EL") based on the complete cases.

[Table 3]

We selected 1000 clustered (two-stage) samples to compute the observed expectation, MSE and coverages of the confidence intervals. In Table 3, we have the observed coverages of the 95% confidence intervals and the tail error rates. The customary empirical likelihood approach and the naïve approach give similar coverages and rates. The naïve approach can give coverages as high as 98.0%, and tail error rate as low as 0.91%. None of the coverages of the empirical likelihood approach proposed are significantly different from the nominal value (95%). The tail error rates are not significantly different from 2.5%, except the right tail error rate of β_2.

8.4 Cluster (two-stage) sampling with large sampling fractions

We consider a cluster sampling design with large sampling fractions. Consider an artificial populations of N clusters with totals values Y_k generated from a skewed distribution given by

$$Y_k = 100 \times \{3 + \zeta_k + \psi(e_k - 1)\},$$

where $\zeta_i \sim \exp(1)$ and $e_i \sim \chi^2_{df=1}$. Here, $\psi = 0.5$ or 2.3. The probabilities π_k are proportional to ζ_k. We have an informative design when $\psi = 0.5$, since the correlation between π_k and Y_k and is approximately 0.8. With $\psi = 2.3$ the correlation is 0.3 and the design is less informative. Let $\widetilde{N}_k \sim \text{Uniform}(200, 500)$. Within clusters $\widetilde{N}_k$ values y_i are generated from the normal distribution

$$y_i \sim N\big(\overline{Y}_k, 0.5\overline{Y}_k\big) \quad \text{for } i \in \widetilde{U}_k,$$

where $\overline{Y}_k := Y_k \widetilde{N}_k^{-1}$. Missing values for y_i are generated as in §8.1.

Two populations are created, by generating two data sets of $N = 2000$ and $N = 12500$ clusters. The resulting size $\mathcal{N}$ of the population $\mathcal{U}$ is approximately 706 000 and 442 500. We select 1000 randomised systematic samples of $n = 500$ clusters. This gives non-negligible sampling fractions 0.25 and 0.4. A simple random sample of 20% of units is selected, within each clusters sampled. Stratification and side information are not used.

The parameters of interest are the quantiles Y_α, with $\alpha = 0.1$, 0.2 and 0.3. The $\boldsymbol{g}_i(\boldsymbol{\tau})$ for quantiles can be found in Berger and Torres (2016 §7.1). Thus, $\boldsymbol{\tau}_0 = \theta_0$, where $\theta_0 = Y_\alpha$. Another parameter of interest is the distribution function of the y_i,

$$F(Y) := \mathcal{N}^{-1} \sum_{i \in \mathcal{U}} \delta\{y_i \leqslant Y\},$$

for a given value Y. Here, $\delta\{y_i \leqslant Y\} = 1$, if $y_i \leqslant Y$ and $\delta\{y_i \leqslant Y\} = 0$ otherwise. In this cases, $\boldsymbol{g}_i(\boldsymbol{\tau}) = \delta\{y_i \leqslant Y\} - \theta$. We consider $F(Y_{0.1})$ and $F(Y_{0.3})$. Thus, $\theta_0 = 0.1$ or 0.3. Here, $\boldsymbol{\tau}_0 = \theta_0$, where $\theta_0 = F(Y_{0.1}) = 0.10$ or $F(Y_{0.3}) = 0.3$

In all cases, we have three unknown parameters: the parameter of interest and two response parameters within $\boldsymbol{\lambda}_0$; that is, $\boldsymbol{\psi}_0 = (\boldsymbol{\tau}_0^\top, \boldsymbol{\lambda}_0^\top)^\top$, with $\boldsymbol{\tau}_0 = \theta_0$ being either a quantile or a distribution function, and $\boldsymbol{\lambda}_0 = (-1, 1)^\top$. The 95% confidence intervals are constructed using (39); that is,

$$\textsc{ci}(\theta_0) := \Big\{\theta : \widehat{R}(\theta)\widehat{\lambda}_1^{-1} \leqslant 3.8415\Big\}. \tag{44}$$

This confidence interval will be compared with the non-adjusted empirical likelihood confidence interval (40) and the standard confidence interval based on linearisation (e.g. Deville 1999) and the central limit theorem with the traditional two-stage variance (e.g. Särndal et al 1992, p137) containing a variance component due to non-response, as in Shao and Steel (1999). The confidence interval (40) and (44) are range preserving; that is, the bound are within the parameter space. Thus, since $0 < \theta_0 < 1$, the lower bound is always larger than 0 and the upper bound is always smaller than 1. Range preserving is not guaranteed with the standard confidence interval.

[Table 4]

The coverages and tail error rates are given in Table 4. The standard linearisation approach suffers from a low coverage, due to a bias in the variance estimator, the lack of normality and the fact that the bounds could be outside the parameter space. The bias and lack of normality can be explained by the skewness of the data. The low coverage of confidence intervals based on linearised variance is a know issue (Valliant 2004; Graf and Tillé 2014). The empirical likelihood approaches gives coverages closer to the nominal value (95%). These coverages seems not to be related to the correlation between π_k and Y_k. The range of the correction $\widehat{\lambda}_1$ is [0.77, 1.07] with an average of 0.93. As a result, the confidence intervals (44) gives coverages slightly smaller than those obtained with (40). With the larger sampling fraction ($N = 1250$), we have a larger difference between the coverages and a smaller correction $\widehat{\lambda}_1$. The simulation suggests that the effect of $\widehat{\lambda}_1$ is small compared to the difference between the coverage and the nominal value. Overall, the non-adjusted confidence interval (40) gives coverages closer to 95%. Thus, the simulation study suggests that the more conservative confidence interval based on (40) seems preferable. More explanation can be found at the end of §7.

9 An application to the educational survey data (PISA)

The empirical likelihood approach proposed is applied to the 2006 PISA survey data (OECD 2006, 2007) for the United Kingdom, containing information on the skills and knowledge of 13 152 fifteen year-old students. This dataset has missing values. A two-stage sampling design was used. The schools are the clusters and the pupils are the units. We use the reciprocal unit level and cluster level weights as proxies for the inclusion probabilities π_k and $\pi_{i|k}$ defined in §3.

We consider the logistic model (43) to explain the probability of a mathematics achievement score below 497.27, which is the mean observed from the data. The vector $\boldsymbol{x}_i$ contains the following explanatory variables

– *Parent-tertiary*: 1 if parents have tertiary education, 0 otherwise
– *Male*: 1 for males, 0 for females
– *Large-class*: 1 for class size over 25, 0 otherwise
– *City*: 1 for city located schools, 0 otherwise

One component of $\boldsymbol{x}_i$ is 1 for the intercept. The response variable y_i and the variables *Parent-tertiary*, *Large-class* and *City* contains missing values.

The side information is the fraction of fifteen year-old males in 2006, which is 51.5%, according to the OECD website: `http://stats.oecd.org`. Thus, $\mathbf{f}_i = Male_i - \boldsymbol{\varphi}_0$, where $\boldsymbol{\varphi}_0 = 0.515$. The males are under-represented in the PISA survey, because the weighted estimates of the proportion of males is 49.5%, which is lower than 51.5%. The side information corrects for this under-representation.

For the non-response mechanism, we consider the following additional variables

– *Scotland*: 1 for schools in Scotland, 0 for schools in England and Wales
– *Public*: 1 for public school, 0 otherwise

The variables *Scotland* and *Public* are cross-classified into four groups. A descriptive analysis reveals that these groups and the variable *Male* are significant to explain non-response. Hence, $\boldsymbol{\xi}_i$ contains the variable *Male*, three dichotomous variables specifying the groups and a variable equal to 1 for the intercept. The baseline group is the public schools in England and Wales. The model (2) is considered with $\digamma$ being the logit function.

The parameter is $\boldsymbol{\psi}_0 = (\boldsymbol{\tau}_0^\top, \boldsymbol{\lambda}_0^\top)^\top$, where $\boldsymbol{\tau}_0$ are the regression coefficients of $\boldsymbol{x}_i$ in the model (43). The response parameter $\boldsymbol{\lambda}_0$ contains the coefficient of $\boldsymbol{\xi}_i$ in the model (2). Hence, $\boldsymbol{\psi}_0 \in \mathbb{R}^{10}$. The pivotal statistics (19) is used to compute the p-values for each component θ_0 of $\boldsymbol{\tau}_0$. Indeed, (36) implies that $\widehat{R}(\theta_0)$ converge to a χ^2-distribution with one degree of freedom, under $H_0 : \theta_0 = 0$.

In Table 5, we have the empirical likelihood estimates and p-values computed from (19) and the p-values of Owen's (2001) customary empirical likelihood approach (Column "*Customary EL*") based on the complete cases. We also have those obtain from the naïve approach which consists in fitting a logistic model from the complete cases, using maximum likelihood. There is no difference between the p-values of the naïve and customary empirical likelihood approach. This is in-line with the coverages and error rates observed in Table 3. However, the empirical likelihood approach proposed give different p-values. *Parent-tertiary* and *Male* are significant in all cases, but with different estimates. The intercept is not significant with the approach proposed. Note that *Large-class* is only significant with the approach proposed, despite that this effect is not significant with the other approaches. The cluster effect and non-response may explain the differences between the p-values of the intercept and *Large-class*. Unlike, the customary empirical likelihood and naïve approaches ignore the weights, the empirical likelihood approach proposed takes into account of the clustering, the weights, the non-response mechanism and the estimation of the response parameter.

Appendix A

In this Appendix, we propose an estimator for (38). We have that (see (C.10) and (C.11) in Appendix C of the online supplement)

$$\boldsymbol{V}_0 = \boldsymbol{V}_0^{\mathrm{I}} + \boldsymbol{V}_0^{\mathrm{II}} + \boldsymbol{o}(1)\cdot \tag{A.1}$$

where

$$\begin{aligned} \boldsymbol{V}_0^{\mathrm{I}} &:= \mathbb{E}_r\{\mathbb{V}_d(\bar{\boldsymbol{\epsilon}}_\pi \mid \boldsymbol{r})\} \\ \boldsymbol{V}_0^{\mathrm{II}} &:= \mathbb{V}_r\{\mathbb{E}_d(\bar{\boldsymbol{\epsilon}}_\pi \mid \boldsymbol{r})\} \\ \bar{\boldsymbol{\epsilon}}_\pi &:= N^{-1} \sum_{k \in \widetilde{\boldsymbol{S}}} \pi_k^{-1} \widehat{\boldsymbol{\epsilon}}_k \end{aligned} \tag{A.2}$$

and $\widehat{\boldsymbol{\epsilon}}_k$ is defined by (29). The operators $\mathbb{E}_r(\cdot)$ and $\mathbb{V}_r(\cdot)$ denote the expectation and variance with respect to the response mechanism. The operators $\mathbb{V}_d(\cdot \mid \boldsymbol{r})$ and $\mathbb{E}_d(\cdot \mid \boldsymbol{r})$ denote the conditional expectation and variance with respect to the sampling design, given $\boldsymbol{r}$. An asymptotically unbiased estimator of $\boldsymbol{V}_0^{\mathrm{I}}$ is

$$\widehat{\boldsymbol{V}}_0^{\mathrm{I}} := \widehat{\mathbb{V}}_d(\bar{\boldsymbol{\epsilon}}_\pi \mid \boldsymbol{r}) \tag{A.3}$$

where $\widehat{\mathbb{V}}_d(\bar{\boldsymbol{\epsilon}}_\pi \mid \boldsymbol{r})$ denotes the customary two-stage variance estimator of $\mathbb{V}_d(\bar{\boldsymbol{\epsilon}}_\pi \mid \boldsymbol{r})$ (e.g. Särndal et al 1992, p137), treating $\boldsymbol{r}$ as constant. This estimator takes into account of large sampling fractions, because it depends on the joint-inclusion probabilities of the clusters. The second term $\boldsymbol{V}_0^{\mathrm{II}}$ can be estimated by (see (C.12) in Appendix C of the online supplement)

$$\widehat{\boldsymbol{V}}_0^{\mathrm{II}} := N^{-2} \sum_{k \in \widetilde{\boldsymbol{S}}} \pi_k^{-1} \sum_{i \in \boldsymbol{S}_k} \pi_{i|k}^{-1} \boldsymbol{\kappa}_i(\boldsymbol{\psi}_0)^\top \boldsymbol{\kappa}_i(\boldsymbol{\psi}_0)\ P_i(\boldsymbol{\lambda}_0)\{1 - P_i(\boldsymbol{\lambda}_0)\}, \tag{A.4}$$

where $P_i(\boldsymbol{\lambda}_0)$ is defined by (3) and

$$\boldsymbol{\kappa}_i(\boldsymbol{\psi}_0) := \{P_i(\boldsymbol{\lambda}_0)^{-1} \boldsymbol{g}_i(\boldsymbol{\tau}_0)^\top, \boldsymbol{\xi}_i^\top\}^\top\cdot \tag{A.5}$$

The unknown quantity $\boldsymbol{\beta}_0$ is substituted by $\widehat{\boldsymbol{b}}_0$ within (A.3) and (A.5).

Finally, (A.1), (A.3) and (A.4) gives the following estimator for (38)

$$\widehat{\boldsymbol{\mathcal{L}}}_0 = (\widehat{\boldsymbol{V}}_0^{\mathrm{I}} + \widehat{\boldsymbol{V}}_0^{\mathrm{II}})(\boldsymbol{I} - \widehat{\boldsymbol{A}}_0)\ \widehat{\boldsymbol{V}}_0^{-1}\cdot \tag{A.6}$$

The estimates $\widehat{\lambda}_1, \dots, \widehat{\lambda}_p$ of $\lambda_1, \dots, \lambda_p$ are the eigenvalues of (A.6), after substituting $\boldsymbol{\psi}_0$ by $\widehat{\boldsymbol{\psi}}$ within the right hand side of(A.6).

Acknowledgements This work was supported by the European Unions's Sevenths Programme for Research, Technological Development and Demonstration under Grant Agreement No 312691 - InGRID. I wish to thanks Dr. Melike Oğuz-Alper (Statistics Norway) for useful comments and help with §9. I also wish to thank an anonymous reviewer for suggesting adding sections 7, 8.4 and 9.

Supplement

The detailed proof of Lemma 2, Lemma 3 and Theorem 1 can be found in the online supplement.

References

Alfons A, Filzmoser P, Hulliger B, Kolb J, Kraft S, Münnich R (2011) Synthetic Data Generation of SILC Data. Research Project Report WP6 – D6.2, University of Trier, `http://ameli.surveystatistics.net`

Berger YG (2018) Empirical likelihood approaches under complex sampling designs. In: Wiley StatsRef: Statistics Reference Online, Wiley, DOI 10.1002/9781118445112.stat08066

Berger YG, Rao JNK (2006) Adjusted jackknife for imputation under unequal probability sampling without replacement. J R Statist Soc B 68:531–547

Berger YG, Torres ODLR (2016) An empirical likelihood approach for inference under complex sampling design. Journal of the Royal Statistical Society Series B 78(2):319–341

Binder DA (1983) On the variance of asymptotically normal estimators from complex surveys. Int Stat Rev 51(427):279–292

Binder DA, Patak Z (1994) Use of estimating functions for estimation from complex surveys. Journal of the American Statistical Association 89(427):1035–1043

Brewer K, Gregoire T (2009) Introduction to survey sampling. In: Pfeffermann D, Rao C (eds) Sample Surveys: Design, Methods and Applications, Handbook of Statistics, Elsevier, Amsterdam, pp 9–38

Brick J, Kalton G (1996) Handling missing data in survey research. Statistical Methods in Medical Research 5:215238

Brick JM, Montaquila JM (2009) Nonresponse and weighting. In: Pfeffermann D, Rao CR (eds) Sample Surveys: Design, Methods and Applications, Handbook of Statistics, vol 29A, Elsevier, Amsterdam, pp 163–185

Chen J, Sitter RR (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. Statist Sinica 9:385–406

Chen S, Kim JK (2014) Population empirical likelihood for nonparametric inference in survey sampling. Statist Sinica 24:335–355

Chen S, Van Keilegom I (2009) A review on empirical likelihood methods for regression. Test 18:415–447

Deville JC (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology 25:193–203

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association 87(418):376–382

Eurostat (2012) European union statistics on income and living conditions (EU-SILC). `http://ec.europa.eu/eurostat/web/income-and-living-conditions/overview`

Fang F, Hong Q, Shao J (2009) A pseudo empirical likelihood approach for stratified samples with nonresponse. Ann Statist 37(1):371–393, DOI 10.1214/07-AOS578, URL `http://dx.doi.org/10.1214/07-AOS578`

Fang F, Hong Q, Shao J (2010) Empirical likelihood estimation for samples with nonignorable nonresponse. Statistica Sinica 20:263–280

Fay BE (1991) A design-based perspective on missing data variance. Proceeding of the 1991 Annual Research Conference US Bureau of the Census pp 429–440

Fuller WA (2009) Some design properties of a rejective sampling procedure. Biometrika 96:933–944

Godambe V, Thompson ME (1974) Estimating equations in the presence of a nuisance parameter. The Annals of Statistics 2(3):568–571

Godambe VP, Thompson M (2009) Estimating functions and survey sampling. In: Pfeffermann D, Rao C (eds) Sample Surveys: Inference and Analysis, Handbook of Statistics, Elsevier, Amsterdam, pp 83–101

Graf E, Tillé Y (2014) Variance estimation using linearization for poverty and social exclusion indicators. Survey Methodology (40):61–79

Hájek J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics 35(4):1491–1523

Hartley HO, Rao JNK (1962) Sampling with unequal probabilities without replacement. Ann math Statist Assoc 33:350–374

Hartley HO, Rao JNK (1968) A new estimation theory for sample surveys. Biometrika 55(3):547–557

Haziza D (2009) Imputation and inference in the presence of missing data. In: Pfeffermann D, Rao CR (eds) Sample Surveys: Design, Methods and Applications, Handbook of Statistics, vol 29A, Elsevier, Amsterdam, pp 215–246

Haziza D, Beaumont JF (2007) On the construction of imputation classes in surveys. International Statistical Review 75(1):25–43

Haziza D, Lesage E (2016) A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics 32:129–145

Imbens GW, Lancaster T (1994) Combining micro and macro data in microeconometric models. The Review of Economic Studies 61(4):655–680

Isaki CT, Fuller WA (1982) Survey design under the regression super-population model. Journal of the American Statistical Association 77:89–96

Kalton G (1983) Compensating for missing survey data. University of Michigan Press, Ann Arbor, MI

Kovar JG, Rao JNK, Wu CFJ (1988) Bootstrap and other methods to measure errors in survey estimates. The Canadian Journal of Statistics 16:25–45

Krewski D, Rao JNK (1981) Inference from stratified sample: properties of linearization jackknife, and balanced repeated replication methods. Annals of Statistics 9:1010–1019

Little R (1986) Survey nonresponse adjustments for estimates of means. International Statistical Review 54:139–157

Little R, Rubin DB (2002) Statistical Analysis With Missing Data, 2nd edn. Wiley, Hoboken, NJ

Little R, Vartivarian S (2005) Models for nonresponse in sample surveys. Survey Methodology 31:161–168

Lundström S, Särndal CE (1999) Calibration as a standard method for treatment of nonresponse. Journal of Official Statistics 15:305327

Montaquila J, Brick J, Hagedorn M, Kennedy C, Keeter S (2008) spects of nonresponse bias in rdd telephone surveys. In: Lepkowski J, Tucker C, Brick J, et al (eds) Advances in Telephone Survey Methodology, Wiley, New York

National Center for Health Statistics (2016) National health and nutrition examination survey (NHANES). http://www.cdc.gov/nchs/nhanes

Neyman J (1938) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97(4):558–625

OECD (2006) PISA 2006 Technical Report. https://www.oecd.org/pisa/data/42025182.pdf

OECD (2007) PISA 2006: Science Competencies for Tomorrows World, Volume 1 - Analysis. OECD Publisher, Paris

Oğuz-Alper M, Berger YG (2016) Empirical likelihood approach for modelling survey data. Biometrika 103(2):447–459

Osier G, Berger YG, Goedemé T (2013) Standard error estimation for the eu-silc indicators of poverty and social exclusion. Eurostat Methodologies and Working Papers series

Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75(2):237–249

Owen AB (2001) Empirical Likelihood. Chapman & Hall, New York

Pfeffermann D, Skinner C, Holmes D, Goldstein H, Rasbash J (1998) Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society Series B 60:23–40

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. Ann Statist 22(1):pp. 300–325

Qin J, Zhang B, Leung DHY (2009) Empirical likelihood in missing data problems. Journal of the American Statistical Association 104(488):1492–1503

R Development Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria

Rao C (1973) Linear Statistical Inference and Its Applications, vol 2nd. ed. John Wiley and Sons, New York

Rao JNK, Shao AJ (1992) Jackknife variance estimation with survey data under hotdeck imputation. Biometrika 79:811–822

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rust K, Rao J (1996) Variance estimation for complex surveys using replication techniques. Biometrika 5(3):281310

Särndal CE, Lundström S (2005) Estimation in Surveys with Nonresponse. Wiley, Chichester

Särndal CE, Swensson B (1987) A general view of estimation for two-phases of selection with applications to two-phase sampling and non-response. International Statistical Review 55:279294

Särndal CE, Swensson B, Wretman J (1992) Model Assisted Survey Sampling. Springer-Verlag, New York

Scheffé H (1959) The Analysis of Variance. John Wiley and Sons, New York

Shao J, Steel P (1999) Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. Journal of the American Statistical Association 94:254–265

Valliant R (2004) The effect of multiple weighting steps on variance estimation. Journal of Official Statistics 20:118

Van Der Vaart AW (1998) Asymptotic Statistics. Cambridge University Press, Cambridge

Wang D, Chen SX (2009) Empirical likelihood for estimating equations with missing values. Annals of Statistics 37(1):490517

Wang Q, Rao JNK (2002a) Empirical likelihood-based inference in linear models with missing data. Scandinavian Journal of Statistics 29:563–576

Wang Q, Rao JNK (2002b) Empirical likelihood-based inference under imputation for missing response data. The Annals of Statistics 30(3):896–924

Wolter KM (2007) Introduction to Variance Estimation, 2nd edn. Springer, New York

Wu C, Rao JNK (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. Canad J Statist 34(3):359–375

Wu C, Zhao P, Haziza D (2017) Empirical likelihood inference for complex surveys and the design-based oracle variable selection theory. In: Proceedings of the Section on Survey Research Methods, American Statistical Association, http://ww2.amstat.org/sections/srms/Proceedings

Table 1 Observed coverages of 95% confidence intervals and (left and right) tail error rates. 'MSE': mean squared errors. Corr(y_i,π_i) denotes the correlation between y_i and π_i. 'EL': empirical likelihood approach. 'stand.': standard approach based on (41) and (42). 10 000 samples.

	Coverages (%)		Left rates (%)		Right rates (%)		MSE	
Corr(y_i,π_i)	EL (40)	stand.	EL (40)	stand.	EL (40)	stand.	EL (40)	stand.
0.02	96.3†	91.0†	2.68	0.71†	0.72†	8.34†	17.83	17.85
0.2	95.7†	92.2†	2.91†	0.97†	1.19†	6.81†	0.57	0.57
0.4	94.0†	93.0†	3.29†	1.54†	2.57	5.46†	0.14	0.14
0.6	94.4†	93.7†	2.99†	2.08†	2.61	4.20†	0.05	0.05
0.7	94.9	94.1†	2.71	1.86†	2.33	4.07†	0.04	0.04

† Coverages (or rates) significantly different from 95% (or 2.5%): p-value ≤ 0.05.

Table 2 Observed coverages of 95% confidence intervals and (left and right) tail error rates. 'EL': empirical likelihood approach. In the 9-th column, we have the MSE of the empirical likelihood point estimator divided by the MSE of the Naïve estimator. Corr(y_i,ρ_i) denotes the correlation between y_i and ρ_i. $39600 \leqslant \mathcal{N} \leqslant 39960$. 10 000 samples.

		Coverages (%)		Left rates (%)		Right rates (%)		MSE(EL)	
σ	ϑ	EL (40)	Naïve	EL (40)	Naïve	EL (40)	Naïve	MSE(Naïve)	Corr(y_i,ρ_i)
0.5	−0.5	94.8	78.2†	2.72	21.72†	2.44	0.05†	0.47	−0.17
	−0.4	94.9	79.3†	2.43	20.63†	2.69	0.10†	0.47	−0.19
	−0.2	94.6	86.9†	2.72	12.92†	2.64	0.16†	0.63	−0.18
	0.005	95.0	93.3†	2.50	5.86†	2.52	0.79†	0.91	0.01
1.0	−0.5	95.1	80.2†	2.48	19.68†	2.42	0.09†	0.50	−0.31
	−0.4	94.9	83.5†	2.66	16.33†	2.42	0.14†	0.55	−0.37
	−0.2	95.3	88.5†	2.25	11.16†	2.41	0.36†	0.67	−0.35
	0.005	95.0	94.3†	2.58	4.87†	2.41	0.84†	0.94	0.01
2.0	−0.5	94.7	89.5†	2.81†	10.04†	2.47	0.45†	0.76	−0.54
	−0.4	95.0	90.6†	2.50	8.83†	2.50	0.53†	0.77	−0.58
	−0.2	95.0	92.2†	2.41	7.15†	2.56	0.62†	0.82	−0.55
	0.005	95.4†	93.5†	2.32	5.52†	2.25	0.94†	0.90	0.03
3.0	−0.5	94.9	94.4†	2.56	4.43†	2.51	1.21†	0.99	−0.48
	−0.4	94.9	94.0†	2.58	4.84†	2.52	1.11†	0.98	−0.58
	−0.2	94.9	94.0†	2.55	5.00†	2.51	0.95†	0.93	−0.64
	0.005	95.0	93.5†	2.55	5.62†	2.45	0.91†	0.90	0.03
4.0	−0.5	94.5†	95.2	2.74	2.68	2.71	2.07†	1.06	−0.31
	−0.4	95.1	95.3	2.42	3.08†	2.45	1.65†	1.02	−0.42
	−0.2	95.3	94.5†	2.41	4.20†	2.31	1.26†	0.97	−0.66
	0.005	95.0	93.9†	2.48	5.27†	2.51	0.82†	0.90	0.05

† Coverages (or rates) significantly different from 95% (or 2.5%): p-value ≤ 0.05.

Table 3 Observed coverages of 95% confidence intervals and (left and right) tail error rates. 'EL': empirical likelihood approach. 1000 samples.

	Coverages (%)			Left rates (%)			Right rates (%)		
	EL (40)	Cust. EL	Naïve	EL (40)	Cust. EL	Naïve	EL (40)	Cust. EL	Naïve
β_0: Intercept	94.9	97.1†	97.1†	2.10	1.50†	1.45†	3.00	1.40†	1.45†
β_1: Age	94.5	98.1†	98.0†	2.80	1.00†	1.09†	2.70	0.90†	0.91†
β_2: Educ	93.8	95.0	95.1	2.40	1.70	1.73	3.80†	3.30	3.18
β_3: Married	94.2	93.8	93.9	2.90	3.70†	3.64†	2.90	2.50	2.45
β_4: Male	94.1	94.5	94.5	3.20	2.80	2.73	2.70	2.70	2.73

† Coverages (or rates) significantly different from 95% (or 2.5%): p-value $\leqslant 0.05$.

Table 4 Observed coverages of 95% confidence intervals and (left and right) tail error rates. 'EL (44)': empirical likelihood approach based upon (44). 'EL (40)': empirical likelihood approach based upon (40). '*Lin.*': linearisation approach. 1000 samples of size $n = 500$. $n/N = 0.4$ with $N = 1250$ and $n/N = 0.25$ with $N = 2000$

			Coverages (%)			Left rates (%)			Right rates (%)		
N	Corr(π_k,Y_k)	θ_0	EL (44)	EL (40)	Lin.	EL (44)	EL (40)	Lin.	EL (44)	EL (40)	Lin.
1250	0.8	$Y_{0.1}$	91.5†	93.1†	95.3	5.45†	3.98†	4.30†	2.91	2.81	0.40†
		$Y_{0.2}$	92.5†	94.0	91.1†	5.58†	4.16†	8.80†	1.81	1.71	0.10†
		$Y_{0.3}$	93.4†	94.8	89.5†	5.07†	3.67†	10.10†	1.41†	1.41†	0.40†
		$F(Y_{0.1})$	91.2†	92.8†	95.0	3.00	2.90	3.80†	5.62†	4.17†	1.20†
		$F(Y_{0.3})$	93.4†	94.7	87.3†	2.70	2.40	5.30†	3.63†	2.64	7.40†
	0.3	$Y_{0.1}$	95.0	96.3	90.7†	3.45	2.26	9.00†	1.30†	1.20†	0.30†
		$Y_{0.2}$	93.9	95.2	88.2†	4.22†	3.20	11.50†	1.61	1.41†	0.30†
		$Y_{0.3}$	92.8†	94.5	90.3†	4.69†	3.30	8.90†	2.41	2.11	0.80†
		$F(Y_{0.1})$	95.2	96.7†	86.0†	1.20†	1.00†	2.20	3.33	2.14	11.80†
		$F(Y_{0.3})$	92.2†	93.5†	86.0†	2.20	2.00	4.90†	5.45†	4.39†	9.10†
2000	0.8	$Y_{0.1}$	93.0†	93.1†	93.6†	4.18†	4.07†	5.80†	2.61	2.61	0.60†
		$Y_{0.2}$	96.1	96.6†	89.7†	2.17	1.95	9.90†	1.51†	1.31†	0.40†
		$Y_{0.3}$	95.3	95.7	88.8†	2.13	1.79	10.60†	2.31	2.21	0.60†
		$F(Y_{0.1})$	93.0†	93.2†	93.5†	2.60	2.60	3.70†	4.27†	4.06†	2.80
		$F(Y_{0.3})$	95.3	95.6	87.0†	2.00	2.00	4.10†	2.48	2.14	8.90†
	0.3	$Y_{0.1}$	93.6†	94.2	85.4†	3.94†	3.44	13.80†	2.01	1.91	0.80†
		$Y_{0.2}$	92.9†	93.6†	86.5†	3.89†	3.53†	12.60†	2.71	2.41	0.90†
		$Y_{0.3}$	91.4†	92.3†	88.2†	5.86†	5.10†	10.60†	2.51	2.41	1.20†
		$F(Y_{0.1})$	93.4†	94.1	83.7†	1.80	1.70	2.60	4.42†	3.81†	13.70†
		$F(Y_{0.3})$	91.8†	92.6†	85.0†	2.70	2.40	5.50†	5.23†	4.79†	9.50†

† Coverages (or rates) significantly different from 95% (or 2.5%): p-value $\leqslant 0.05$.

Table 5 Estimates and p-values of the logistic regression based upon the 2006 PISA survey data (OECD 2006, 2007) for the United Kingdom.

	Empirical likelihood		Customary EL		Naïve	
	Estimates	p-value	Estimates	p-value	Estimates	p-value
Intercept	0.13	0.257	0.28	< 0.001†	0.28	< 0.001†
City	0.13	0.443	0.06	0.159	0.06	0.159
Large-class	0.31	0.010†	0.03	0.477	0.03	0.477
Parent-tertiary	−0.52	< 0.001†	−0.49	< 0.001†	−0.49	< 0.001†
Male	−0.46	< 0.001†	−0.32	< 0.001†	−0.32	< 0.001†

† p-value $\leqslant 0.01$.